Πανεπιστήμιο Αιγαίου

Σχολή Επιστημών της Διοίκησης

Τμήμα Μηχανικών Οικονομίας και Διοίκησης

Θεματική ομαδοποίηση αναρτήσεων σε κοινωνικά δίκτυα

Τίνα Ίβκοβα

Τριμελής Επιτροπή:

Ν. Αμπαζής, Επιβλέπων Καθηγητής

Δ. Δριβαλιάρης, Μέλος Επιτροπής

Ε. Βασιλείου, Μέλος Επιτροπής

Χίος, 2017

Ευχαριστίες

Με κάθε ειλικρίνεια, θέλω να εκφράσω την τεράστια ευγνωμοσύνη μου στον επιβλέποντα καθηγητή μου, κ. Νικόλαο Αμπαζή, για την ατελείωτη υπομονή που έδειξε καθ' όλη την διάρκεια της συνεργασίας μας, καθώς και για τις πολύτιμες ώρες που μου αφιέρωσε. Ο τρόπος με τον οποίον με αντιμετώπιζε, άλλοτε με απόλυτη αυστηρότητα, και άλλοτε με μεγάλη επιείκεια, με ενέπνεε να συνεχίζω και να μην απογοητεύομαι στις δυσκολίες που μου έτυχαν κατά την συγγραφή της παρούσας διπλωματικής μελέτης. Όσες γνώσεις και εμπειρίες απέκτησα στα πέντε χρόνια της φοίτησής μου στην σχολή, άλλα τόσα και περισσότερα έχω κερδίσει από την συνεργασία με τον κ. Αμπαζή και τον ευχαριστώ πολύ για όλα.

Θα ήθελα να ευχαριστήσω τον κ. Δριβαλιάρη για την πολύτιμη συνεισφορά του στην παρούσα διπλωματική μου εργασία, αλλά κυρίως για τον συνεχή αγώνα που δίνει όλα αυτά τα χρόνια για να κάνει όλους εμάς να κατανοήσουμε λίγο περισσότερα τα μαθηματικά και να έχουμε για κίνητρο μια δουλειά που να αγαπάμε, όσο αγαπάει αυτός την δική του.

Επίσης, θέλω να ευχαριστήσω τον Γ. Μελέκο για την τεχνική υποστήριξη, χωρίς την οποία δεν θα μπορούσα να ολοκληρώσω αυτήν την διπλωματική μελέτη. Λόγω έλλειψης εμπειρίας είχαν προκύψει προβλήματα τα οποία δεν θα μπορούσα να τα αντιμετωπίσω μόνη μου και είμαι πολύ ευγνώμων για όλη την βοήθεια και τον χρόνο που μου είχε αφιερωθεί όποτε το χρειαζόμουν.

Τέλος, ένα μεγάλο ευχαριστώ στην οικογένεια μου, που μου προσέφερε όλα τα εφόδια για να σπουδάσω. Είμαι ευγνώμων σε πολλούς καθηγητές που γνώρισα εδώ, για τις γνώσεις, καθώς και για τις συμβουλές που μοιράστηκαν μαζί μας, είναι κάτι το οποίο θα με ακολουθεί μια ζωή.

Περιεχόμενα

Περίληψη

Τα κοινωνικά δίκτυα έχουν ενσωματωθεί στην καθημερινότητα των περισσότερων ανθρώπων, με αποτέλεσμα να αποτελούν ένα ισχυρό μέσο ενημέρωσης. Στα πλαίσια της συγκεκριμένης διπλωματικής εργασίας έχει επιλεγεί ο δημοφιλέστερος ιστότοποτος της Ρωσίας για την ανάδειξη του τουριστικού ενδιαφέροντος ως προς την Ελλάδα. Έχει διαπιστωθεί πως το VK αποτελεί έναν σύγχρονο τρόπο επικοινωνίας και ανταλλαγής απόψεων σε μία μεγάλη και ισχυρή χώρα όπως η Ρωσία.

Για την λήψη των επιθυμητών δεδομένων έγινε έρευνα των σχετικών ως προς την Ελλάδα ομάδων στο VK για την μετέπειτα επιλογή των πιο πολυσύχναστων σελίδων. Τα δεδομένα αντλούνται με την βοήθεια κατάλληλων μεθόδων του API που προσφέρει το VK και στην συνέχεια δημιουργούν ένα πίνακα που περιέχει το εξεταζόμενο corpus. Οι γραμμές του πίνακα αποτελούνται από δημοσιεύσεις και σχόλια που έχουν αναρτηθεί από τους χρήστες αυτών των σελίδων. Έπειτα ακολουθεί η διαδικασία της μαθηματικής αναπαράστασης των κειμένων, έτσι ώστε να είναι εφικτή η θεματική τους ομαδοποίηση.

Η διαδικασία μετατροπής των κειμένων σε διανύσματα πραγματοποιείται με την κατανεμημένη αναπαράσταση της φυσικής γλώσσας, στο τέλος της οποίας οι λέξεις αντιστοιχούν σε τέτοια σημεία του διανυσματικού χώρου, ώστε οι αποστάσεις μεταξύ των σημείων αυτών να αντικατοπτρίζουν την σημασιολογική σχέση μεταξύ τους. Δηλαδή, στο τέλος της διαδικασίας, είναι δυνατόν να προσθαφαιρούνται οι λέξεις, ενώ σαν αποτέλεσμα της πράξης προκύπτει μια λίστα από τις πιο σχετικές λέξεις.

Η κατανεμημένη αναπαράσταση στην επεξεργασία της φυσικής γλώσσας επιτυγχάνεται με δύο αλγόριθμους: Word2vec και Doc2vec. Ο πρώτος μετασχηματίζει την κάθε λέξη σε διάνυσμα με τέτοιο τρόπο, ώστε τα σημεία που αντιστοιχούν στις λέξεις με όμοιες έννοιες να βρίσκονται κοντά, ενώ τα διανύσματα των λέξεων που δεν έχουν σχέση μεταξύ τους, να απέχουν πιο πολύ. Έτσι οι αποστάσεις στον διανυσματικό χώρο αντικατοπτρίζουν τις συσχετίσεις των λέξεων. Από την άλλη, ο αλγόριθμος Doc2vec είναι η εξελιγμένη μορφή του αλγορίθμου Word2vec. Δηλαδή όσα ισχύουν στον πρώτο, ισχύουν και στον δεύτερο, με την διαφορά ότι γίνεται ανάθεση διανυσμάτων όχι μόνο στις λέξεις, αλλά και στο ίδιο το

κείμενο, αυξάνοντας κατά αυτόν τον τρόπο και άλλο την εγκυρότητα των αποτελεσμάτων.

Στην συνέχεια, τα κείμενα εκφρασμένα με μαθηματική αναπαράσταση, μπορούν ή να ομαδοποιηθούν, είτε να ταξινομηθούν. Στην προκειμένη περίπτωση, η απουσία ετικετών για κάθε κείμενο μας οδηγεί στην επιλογή της ομαδοποίησης. Αφού ομαδοποιηθούν τα δεδομένα, μπορεί να γίνει διαχώρισμός των θεμάτων που εμπεριέχονται στις ομάδες που έχουν προκύψει. Στα πειραματικά αποτελέσματα διακρίνονται τα θέματα που απασχολούν τους ρωσόφωνους χρήστες του κοινωνικού δικτύου και τι συζητιέται πιο συχνά, όπως οι τουριστικές προτιμήσεις, οι πολιτικές εξελίξεις ή τα θρησκευτικά αξιοθέατα.

Λέξεις κλειδιά: : VK, API, κατανεμημένη αναπαράσταση, επεξεργασία φυσικής γλώσσας (ΕΦΓ), Word2vec, Doc2vec, ομαδοποίηση, LDA

Abstract

Social media have been integrated into most people's lives, thus they've become powerful information providers. In the current diploma thesis we've chosen the most popular Russian website (VK) to analyze the Russian visitors' interest for Greece. It's been known that «Vkontakte» is a major hub of communication and exchange of opinion in Russia.

In order to select the data the most relevant VK pages related to Greece were identified. After the gathering of data by API methods was created a matrix which contained the whole vocabulary. The rows of matrix consisted of the post's texts, so that clustering would be possible through a numerical representation of texts.

The process of text representation as a fixed-length vector was achieved by distributed representation of the natural language, at the end of which, the words were mapped in a vector space in such a way, so as that the semantic meaning of words, was correlated with the distance between the vectors. For example after a model was trained, the words could be treated with simple vector algebra, as if they were numbers and the result was an output list of similar words.

The distributed representation of the natural language processing utilized two algorithms: Word2vec and Doc2vec. The first algorithm converts each word to a vector so that the words with similar meaning are close to each other in the vector space, while other words are more distant. Thus the distances in vector space reflect the semantics of words. Doc2vec algorithm is an extension of Word2vec, where paragraphs are treated as words.

As soon as the documents have been converted into a numerical representation, clustering or classification is possible. Since the posts that have been selected have no labels, clustering (e.g. with k-means) is the only possible choice. Each cluster can then be segregated into several topics which can identify. The interests of Russian users in topics like politics or religion landmarks, and others.

Keywords: VK, API, distributed representation, Natural Language Processing (NLP), Word2vec, Doc2vec, K-Means, LDA

Λίστα Πινάκων

Λίστα Σχημάτων

Επεξήγηση όρων

ΟΥΑ: Ομοσπονδιακή Υπηρεσία Ασφάλειας

ΠΘΜ: Πιθανολογικό Θεματικό Μοντέλο

ΡΔΡ: Ρωσική Δημόσια Ραδιοτηλεοπτική

API: Application Programming Interface

CBOW: Continuous Bag of Words

DM: Distributed Memory

DBOW: Distributed Bag of Words

HMM: Hidden Markov Model

ID: Identification

LDA: Latent Dirichlet Allocation

LTHM: Latent Topic Hypertext Model

LDAvis: Latent Dirichlet Allocation Visualization

PV: Paragraph Vector

PCA: Principal Component Analysis

TF-IDF: Term-Frequency Inverse-Document-Frequency

VK: VKontakte

1. Εισαγωγή

Σημαντικότητα του Τουρισμού

Η δυσμενής κατάσταση που επικρατεί τα τελευταία χρόνια στην Ελλάδα οδηγεί τους επιχειρηματίες να επενδύουν στους πιο κερδοφόρους τομείς της οικονομίας για να ελαχιστοποιήσουν την πιθανότητα αποτυχίας. Ένας από τους σημαντικότερους κλάδους που συνεχίζει να συμβάλλει σημαντικά στην οικονομία της χώρας είναι ο τουρισμός. Τα ανταγωνιστικά πλεονεκτήματα είναι αρκετά ώστε να επικεντρωθεί η αγορά στον συγκεκριμένο κλάδο [22].

Σε αυτήν την διπλωματική εργασία δίνεται έμφαση στις ανάγκες των Ρώσων τουριστών για διάφορους λόγους. Ο σημαντικότερος ίσως αφορά το γεγονός ότι είναι μια τεράστια σε έκταση χώρα η οποία κατοικείται από περίπου 146 εκ. [https://en.wikipedia.org/wiki/Russia] πολίτες και πιθανούς τουρίστες, που θα ήθελε ο κάθε προορισμός να προσελκύσει. Υπάρχει μεγάλη ανάγκη από την άποψη της ανάπτυξης της οικονομίας, να διατηρούνται οι μεγάλες και σημαντικές αγορές, όπως η Κίνα, η Γερμανία, η Τουρκία, το Ισραήλ, και φυσικά ανάμεσα σε αυτές είναι και η Ρωσία.

Κάθε χρόνο γίνονται προσπάθειες από την κυβέρνηση και τις επιχειρήσεις για αναβάθμιση των τουριστικών υπηρεσιών. Πολύ σημαντική υπήρξε η δημιουργία των απευθείας αεροπορικών συνδέσεων μεταξύ της Μόσχας και της Αγ. Πετρούπολης με διάφορους ελληνικούς προορισμούς. Όμως φαίνεται πως δεν είναι αρκετό, ενώ λόγω των πολιτικών εξελίξεων στην Ρωσία, το 2016 είχαν κλείσει οι κλασικοί προορισμοί όπως η Αίγυπτος και η Τουρκία. Συγκεκριμένα οι τουρκικές αρχές είχαν δηλώσει απώλειες τάξης 7,6 εκ. Ρώσων τουριστών [23]. Όλο αυτό το πλήθος αναμενόταν να επαναδιοχετευτεί στις μεσογειακές χώρες. Η Ελλάδα όμως, ιδίως τα βορειοανατολικά νησιά, έχασαν το ανταγωνιστικό πλεονέκτημα λόγω της μεταναστευτικής κρίσης.

Παρόλα αυτά, μόνο το 2016 την Ελλάδα επισκέφτηκαν σχεδόν ένα εκατομμύριο τουρίστες από την Ρωσία [9], που σημαίνει αύξηση της τάξης του 30% σε σχέση με την προηγούμενη χρονιά [6]. Η Κρήτη και η Χαλκιδική είναι από τους πιο αγαπημένους προορισμούς τους. Για να διατηρηθεί και να αυξηθεί η τόσο μεγάλη προσέλευση, η ποιότητα εξυπηρέτησης δεν φτάνει να παραμένει σε υψηλό επίπεδο, αλλά χρειάζεται να ληφθούν υπόψη οι τυχόν δυσαρέσκειες των εκάστοτε επισκεπτών. Για παράδειγμα, οι Ρώσοι δεν φημίζονται για την ευχέρειά τους στην Αγγλική γλώσσα, κάνοντας την επικοινωνία με το προσωπικό των ξενοδοχείων, εστιατορίων,

καταστημάτων ιδιαίτερα δύσκολη μερικές φορές. Η επίλυση τέτοιου είδους προβλημάτων μπορεί να ενισχύσει το πολυπόθητο ανταγωνιστικό πλεονέκτημα.

Το κύριο στοιχείο που ενώνει την Ρωσία με την Ελλάδα είναι η κοινή θρησκεία. Και οι δύο λαοί χαρακτηρίζονται για την βαθιά πνευματικότητα και αγάπη για την Ορθοδοξία. Τα τελευταία χρόνια οι ταξιδιωτικοί οδηγοί άρχισαν να εκμεταλλεύονται την επιθυμία των τουριστών για εξερεύνηση των ιερών προορισμών, μετατρέποντας τον εναλλακτικό τουρισμό σε μια νέα εξεζητημένη τάση.

Ο τουριστικός κλάδος προσφέρει σημαντικές επενδυτικές ευκαιρίες που αφορούν κυρίως την ανάπτυξη εξειδικευμένων τουριστικών προϊόντων στον τομέα της γαστρονομίας, πολιτισμού και κάθε είδους πολυτέλειας. Το μέγεθος της συνεισφοράς του τουρισμού στην οικονομία φαίνεται από τον παρακάτω πίνακα [8]:

Πίνακας 1.1 Βασικά Μεγέθη

Συνολική Προστιθέμενη Αξία του Τουρισμού	€28 δισ.[1]
Τουρισμός ως ποσοστό της Ελληνικής Ακαθάριστης Προστιθέμενης Αξίας	16%[2]
Διεθνείς Αφίξεις το 2014	~ 22εκατ.[3]
Θέσεις Απασχόλησης στον Κλάδο Τουρισμού	657.000[4]
Αριθμός παραλιών και μαρίνων με Γαλάζιες Σημαίες	393 Παραλίες 9 Μαρίνες
Αριθμός μαρίνων σε λειτουργία	19[5]
Αριθμός θέσεων ελλιμενισμού	6,661[6]
Αριθμός μνημείων αναγνωρισμένων από την UNESCO ως χώροι Παγκόσμιας Πολιτιστικής Κληρονομιάς	17[7]

1. McKinsey report for Tourism, 2013
2. McKinsey report for Tourism, 2013
3. ΣΕΤΕ, ΤΡΑΠΕΖΑ ΕΛΛΑΔΟΣ
4. ΣΕΤΕ, 2014
5. Ιστότοπος ΕΟΤ
6. Ιστότοπος ΕΟΤ
7. http://whc.unesco.org/en/statesparties/GR/

Τουρισμός & Κοινωνικά Δίκτυα

Οι επιχειρήσεις που επιμένουν να λειτουργούν με τον παραδοσιακό τρόπο και αρνούνται να συμβαδίσουν με την εξέλιξη της τεχνολογίας τις περισσότερες φορές χάνουν το ανταγωνιστικό τους πλεονέκτημα. Για τον λόγο αυτό είναι άκρως σημαντική η υιοθέτηση καινοτομιών από τον τουριστικό τομέα για να αυξηθεί όσο το δυνατόν περισσότερο το μερίδιο αγοράς.

Τα κοινωνικά δίκτυα προσφέρουν άμεση και γρήγορη ενημέρωση και χωρίς ιδιαίτερη δυσκολία έχουν "ριζώσει" ολοκληρωτικά στην καθημερινότητα των περισσότερων ανθρώπων. Οι καταναλωτές συχνά αφήνουν κρητικές και σχόλια για προϊόντα και υπηρεσίες, κάτι το οποίο είναι χρήσιμο στις επιχειρήσεις για την βελτίωση της ποιότητας των αγαθών που προσφέρουν. Έτσι και ο τουριστικός τομέας επενδύει ενεργά τον τελευταίο καιρό στα μέσα κοινωνικής δικτύωσης με τελικό στόχο να προσελκύσει όσο το δυνατόν περισσότερους ενδιαφερόμενους.

Αξίζει να αναφερθεί πως τα κοινωνικά δίκτυα αποτελούν ένα σημαντικό μέσο του μάρκετινγκ, ως ένας εύκολος τρόπος προώθησης και προβολής των θετικών στοιχείων. Συχνά οι χρήστες προτείνουν ο ένας στον άλλον την σελίδα που τους άρεσε, βοηθώντας να εξαπλωθεί η διαφήμιση ακόμα γρηγορότερα. Φυσικά ο τουρισμός δεν αποτελεί την εξαίρεση αυτής της τάσης, καθώς τα ξενοδοχεία, οι τουριστικοί πράκτορες και τα μαγαζιά εστίασης προσπαθούν να χειριστούν έξυπνα την νέα τεχνολογία για να μην χάσουν την ανταγωνιστικότητά τους.

Για να καταλάβει όμως μια επιχείρηση προς ποια κατεύθυνση πρέπει να εξελιχθεί, χρειάζεται να γνωρίζει τις αδυναμίες της για να μπορεί τις διορθώσει. Φυσικά η πηγή όλων των πληροφοριών είναι τα μέσα κοινωνικής δικτύωσης, όπου οι χρήστες αφήνουν τα σχόλιά τους, τις κριτικές και τα παράπονα τους. Βασιζόμενοι σε αυτά τα δεδομένα είναι εύκολο να εντοπιστούν τα αδύναμα σημεία, πάντα με τελικό σκοπό την ικανοποίηση των πελατών.

Σημαντικότητα Δεδομένων

Για την επίτευξη της αύξησης των εσόδων ο τουριστικός τομέας χρειάζεται να προσελκύσει ένα μεγαλύτερο μερίδιο αγοράς. Αυτό απαιτεί γνώση της τωρινής κατάστασης, δηλαδή ποιες είναι οι ανάγκες και προτιμήσεις των τουριστών, τι παράπονα έχουν εκφράσει εκείνοι που έχουν επισκεφτεί αυτήν την χώρα, τι προσφέρουν άλλοι τουριστικοί προορισμοί και τους προτιμάνε περισσότερο από την

Ελλάδα. Απάντηση σε όλα αυτά τα ερωτήματα έρχονται να δώσουν τα δεδομένα από τις κατάλληλες σελίδες των κοινωνικών δικτύων.

Μέσω Application Programming Interfaces (APIs) που προσφέρουν τα περισσότερα κοινωνικά δίκτυα γίνεται η σάρωση των δημοσιεύσεων και όλων των απαραίτητων πληροφοριών που μπορούν να φανούν χρήσιμες. Με την βοήθεια αυτού του εργαλείου είναι εφικτή η συλλογή μεγάλου όγκου δεδομένων. Όσο περισσότερα είναι τα δεδομένα, τόσο προφανώς πιο έγκυρο είναι το συμπέρασμα. Ύστερα από αυτό, η επιχείρηση μπορεί να προβεί στο πρακτικό κομμάτι των απαιτούμενων αλλαγών [29].

Σκοπός και στόχοι της εργασίας

Στην παρούσα διπλωματική μελέτη εφαρμόζεται ο συνδυασμός των αλγορίθμων Word2vec, Doc2vec, K-Means και LDA πάνω στις δημοσιεύσεις που συλλέχθηκαν από το ρωσικό κοινωνικό δίκτυο. Ο στόχος είναι η βέλτιστη εκτίμηση των θεμάτων που απασχολούν τους χρήστες. Για να γίνει εφικτή η ομαδοποίηση των κειμένων, είναι αναγκαία η μετατροπή τους σε μαθηματική αναπαράσταση. Οι δημοσιεύσεις διασπώνται σε λέξεις, οι οποίες με την σειρά τους μετασχηματίζονται σε διανύσματα με τέτοιο τρόπο, ώστε οι αποστάσεις μεταξύ αυτών να αντιπροσωπεύουν την σημασιολογία τους. Η ομαδοποίηση αυτών των διανυσμάτων στην πραγματικότητα είναι η ομαδοποίηση των αρχικών κειμένων, μέσα στις ομάδες των οποίων στην συνέχεια εντοπίζονται τα θέματα.

Δομή της διπλωματικής εργασίας

Η διπλωματική εργασία εστιάζει στην θεματική ανάλυση δεδομένων στα κοινωνικά δίκτυα. Ύστερα από την εισαγωγή, όπου επισημαίνεται η χρησιμότητα της ανάλυσης στον τουριστικό τομέα, ακολουθούν τέσσερα βασικά κεφάλαια. Στο πρώτο παρουσιάζεται το πιο γνωστό ρωσικό κοινωνικό δίκτυο και εξηγείται γιατί επιλέχθηκε το συγκεκριμένο για την λήψη δεδομένων.

Το δεύτερο κεφάλαιο εστιάζει στην μέθοδο της κατανεμημένης αναπαράστασης στην επεξεργασία της φυσικής γλώσσας, όπου παρουσιάζονται δύο τρόποι μετασχηματισμού των κειμένων σε μαθηματική αναπαράσταση στον διανυσματικό χώρο. Τα συλλεγόμενα δεδομένα περνάνε από το μοντέλο «Doc2vec», το οποίο έχει

επιλεγεί ως το καταλληλότερο και στην συνέχεια, το μοντέλο αυτό κρίνεται για την εγκυρότητά του από τα αποτελέσματα που προκύπτουν σε σχέση με τυχαίες λέξεις.

Στο τρίτο κεφάλαιο εφαρμόζεται ο αλγόριθμος «K-Means» για την ομαδοποίηση των δεδομένων. Στα πειραματικά αποτελέσματα διακρίνεται η ομοιομορφία της ομαδοποίησης, κάτι που προδίδει την σωστή επιλογή του αριθμού $\boldsymbol{k}$. Στην συνέχεια παρουσιάζεται η θεματική μοντελοποίηση μέσα σε κάθε ομάδα με βάση την μεθοδολογία Latent Dirichlet Allocation (LDA). Με αυτόν τον τρόπο η κάθε ομάδα χωρίζεται σε επιμέρους θέματα από τα οποία αποτελείται, ενώ με τα εργαλεία που αναλύονται στο τέταρτο κεφάλαιο, πραγματοποιείται η απεικόνιση των αποτελεσμάτων αυτών.

2. Η τάση των Κοινωνικών Δικτύων

Ο όγκος των παραγόμενων δεδομένων στο διαδίκτυο αυξάνεται καθημερινά. Οι χρήστες αναρτούν υλικό, αφήνουν σχόλια και φυσικά χρησιμοποιούν τα κοινωνικά δίκτυα. Σύμφωνα με το Εθνικό ρωσικό κέντρο έρευνας της κοινωνικής γνώμης τα δύο τρίτα των πολιτών της Ρωσίας (70%) χρησιμοποιούν το διαδίκτυο. Τα τελευταία τρία χρόνια αυτό το ποσοστό δεν έχει αλλάξει, όμως η καθημερινή χρήση του ίντερνετ έχει αυξηθεί στο 55% (από 5% το 2006) [20].

Το πιο δημοφιλές κοινωνικό δίκτυο στην Ρωσία εδώ και πολλά χρόνια παραμένει το VK. Το επιλέγουν το 52% των χρηστών του διαδικτύου, ενώ η προτίμησή του από την γενιά των 18 έως 24 χρονών φτάνει την πλειοψηφία των 86%, την στιγμή που οι άνθρωποι των 60 και άνω χρησιμοποιούν την σελίδα «Odnoklasniki» (Συμμαθητές) η οποία χάνει την δημοτικότητά της χρόνο με τον χρόνο [25]. Αξίζει να σημειωθεί πως παρατηρείται μεγαλύτερη προσέλευση στο «Odnoklasniki» από χρήστες με κατάρτιση της πρωτοβάθμιας και δευτεροβάθμιας εκπαίδευσης (50% και 51% αντίστοιχα) και γυναίκες (53%). Τα πιο δημοφιλή κοινωνικά δίκτυα στην Ελλάδα όπως το Whatsapp, Facebook και Instagram δεν έχουν ιδιαίτερη απήχηση στην Ρωσία (18%, 13% και 12% αντίστοιχα) [20].

Στις 26-27 Μαρτίου του 2016 το Εθνικό ρωσικό κέντρο έρευνας της κοινωνικής γνώμης πραγματοποίησε έρευνα στους πολίτες άνω των 18 ετών. Το σφάλμα των αποτελεσμάτων δεν ξεπερνάει το 3,5% που δίνει την δυνατότητα εξαγωγής βάσιμων συμπερασμάτων. Η δημοσκόπηση έδειξε ότι:

Πίνακας 0.1 Ερωτηματολόγιο
Χρησιμοποιείτε το διαδίκτυο, αν ναι, πόσο συχνά ; (κλειστή ερώτηση, μια απάντηση,%)

	2008	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
Σχεδόν Καθημερινά	15	19	23	30	38	42	47	51	**53**
Όχι συχνά (λίγες φορές την εβδομάδα)	18	21	20	20	20	21	21	19	**16**
Σπάνια, όχι λιγότερο από μια φορά το χρόνο	8	5	3	3	2	2	2	2	**1**

Δεν χρησιμοποιώ	59	55	53	47	40	34	30	28	**29**
Δυσκολεύομαι να απαντήσω	0	0	1	0	0	1	0	0	**1**

Ιδιαίτερο ενδιαφέρον αποτελούν τα κοινωνικά δίκτυα, αφού περιέχουν μεγάλο όγκο πληροφοριών που είναι πολύ χρήσιμες για εταιρείες. Είναι σημαντικό να γνωρίζουν την γνώμη των πελατών για ένα καινούργιο προϊόν ή μία νέα είδηση. Επίσης οι προτιμήσεις των χρηστών είναι καθοριστικές για τους χορηγούς, διότι είναι η κύρια πηγή εισοδήματος για τις σελίδες του Internet. Έτσι είναι σημαντικό για παράδειγμα το τμήμα προώθησης να γνωρίζει σε ποιο ιστότοπο θα έχει μεγάλη απήχηση η διαφήμιση που σχεδιάζει να προτείνει. Την επιλογή της πηγής των δεδομένων για ανάλυση στην παρούσα διπλωματική εργασία την καθόρισε η ίδια έρευνα στην οποία φάνηκε ξεκάθαρα ποιο κοινωνικό δίκτυο θα παρέχει πιο ολοκληρωμένο πόρισμα:

Πίνακας 0.2 Επισκεψιμότητα κοινωνικών δικτύων
Επισκέπτεστε τα κοινωνικά δίκτυα? Αν ναι, τότε ποια κοινωνικά δίκτυα χρησιμοποιείτε?(κλειστή ερώτηση, μία ή παραπάνω απαντήσεις, %)

	2012	**2015**	**2016**
Vkontakte	51	55	52
Odnoklassniki	61	54	42
Whatsapp	0	0	18
Facebook	15	16	13
Instagram	0	8	12
Blogs@mail.ru	26	19	10
Twitter	8	7	7
LinkedIn	0	1	0
MySpace	1	1	0

2.1 Τι είναι το VKontakte (VK)

Το VKontakte είναι το πιο διαδεδομένο κοινωνικό δίκτυο στο ρωσόφωνο κοινό με καθημερινή επισκεψιμότητα που ανέρχεται στα 43 εκατομμύρια [12]. Το VK δημιουργήθηκε το 2006 έχοντας την έδρα του στην Αγία Πετρούπολη . Ο ιδρυτής, Πάβελ Ντούροβ, ήταν απόφοιτος φιλολογικού τμήματος του πανεπιστημίου της Αγίας Πετρούπολης μαζί με τον αδερφό του, τον Νικολάι, ο οποίος εξειδικευόταν στον μαθηματικό και τον προγραμματιστικό τομέα. Αν και το έργο αυτό δεν ήταν το πρώτο στο βιογραφικό του Ντούροβ, μιας και είχαν προηγηθεί οι σελίδες όπως durov.com που έμοιαζαν με την ηλεκτρονική βιβλιοθήκη με σημειώσεις για τα φιλολογικά τμήματα, και αργότερα το φόρουμ spbgu.ru για την επικοινωνία μεταξύ των φοιτητών. Από το ξεκίνημα της ύπαρξης αυτών των ιστοσελίδων, ο Πάβελ Ντούροβ είχε εκφράσει την δυσανασχέτηση του σχετικά με την απόκρυψη πραγματικών στοιχείων των μελών της ομάδας, όπως το πραγματικό όνομα ή το πανεπιστήμιο και το τμήμα στο οποίο φοιτούσαν.

Ως βάση για ένα νέο φοιτητικό περιβάλλον όπου ο χρήστης θα έπρεπε να συμπληρώσει τα προσωπικά του στοιχεία και να ανεβάσει την φωτογραφία του, είχε παρθεί το πρότυπο του αμερικανικού κοινωνικού δικτύου Facebook, το οποίο τότε δεν υπήρχε σε ρωσική εκδοχή. Το 2006 έγινε η πρώτη ιδιωτική αξιολόγηση του ιστότοπου που σχεδίαζαν να ονομάσουν students.ru μιας και στην αρχή, ως γνήσια αντιγραφή του Facebook, το VKontakte απευθυνόταν μόνο στους φοιτητές των ακριβοπληρωμένων ανώτατων εκπαιδευτικών ιδρυμάτων, οι οποίοι με ιδιωτική πρόσκληση και ύστερα από έλεγχο και ταυτοποίηση στοιχείων γίνονταν δεκτοί στην διαδικτυακή ελίτ. Λίγο αργότερα έγιναν δεκτά τα μέλη πιο «απλών» πανεπιστημίων, ώσπου η ιστοσελίδα σταμάτησε να είναι το προνόμιο αποκλειστικά των φοιτητών.

Στις 1 Οκτωβρίου του 2006 καταχωρήθηκε η ηλεκτρονική διεύθυνση vkontakte.ru και μετά από ένα μήνα είχε ανοίξει τις εγγραφές. Ήδη στα τέλη του 2006 το VK παρείχε εξεζητημένες λειτουργείες όπως η αναζήτηση των παλιών φίλων και συμμαθητών, το σύστημα προσωπικών μηνυμάτων, αλλά και τη συλλογή φωτογραφιών. Για την προώθηση της καινούργιας κοινωνικής δικτύωσης οι ιδρυτές ανακοίνωσαν έναν διαγωνισμό, οι νικητές του οποίου θα αποκτούσαν ακριβά δώρα.

Παρόλα αυτά το vkontakte δεν ήταν το πρώτο κοινωνικό δίκτυο στην Ρωσία. Την άνοιξη του 2006 ξεκίνησε την πορεία του και το ok.ru (Συμμαθητές) που κατείχε παρόμοια λειτουργικότητα. Κατά την έναρξη της λειτουργίας του VK, το ήδη υπάρχον δίκτυο μετρούσε πάνω από ένα εκατομμύριο χρήστες. Έτσι τα μέσα μαζικής ενημέρωσης παρακολουθούσαν στενά τις εξελίξεις της πορείας των δύο ανταγωνιστών. Όλοι ανέμεναν για το ποιος θα κατακτούσε την πρωτιά των προτιμήσεων των Ρώσων χρηστών. Τον Ιούλιο του 2007 το VK ήταν στην πρώτη δεκάδα των πιο δημοφιλών ιστότοπων του Ρωσικού δικτύου «Ρουνέτ», καθώς τότε είχε εγγραφεί ο εκατομμυριοστός χρήστης. Ύστερα από έναν χρόνο, όπως ανακοίνωσε ο Ντούροβ, το VK ήταν μια από τις πιο επισκέψιμες σελίδες στην Ρωσία, συγκεκριμένα το 2009, όταν ξεπέρασε τους «Συμμαθητές» η ανάπτυξη του οποίου έπεσε λόγω της κατάργησης της δωρεάν εγγραφής. Όταν το 2010 ο αριθμώς των χρηστών ξεπέρασε τα 70 εκατομμύρια, το VK έφτασε στην τρίτη θέση του ρωσικού δικτύου ακολουθώντας το Yandex και το Mail.ru.

Τον Φεβρουάριο του 2011 η διοίκηση του VK κατήργησε την δωρεάν εγγραφή των χρηστών. Πλέον η δημιουργία καινούργιου λογαριασμού θα ήταν εφικτή μόνο μέσω πρόσκλησης από έναν ήδη υπάρχοντα χρήστη. Το εμπόδιο αυτό είχε στόχο την καταπολέμηση των διαφημιστικών ρομπότ. Βέβαια αυτά τα μέτρα έπαψαν να ισχύουν τον Ιούλιο της ίδιας χρονιάς. Παρά την δημοτικότητα του, το VK δεν είχε βραβευτεί ποτέ από τον οργανισμό Ρουνέτ, μονάχα το 2007 έφτασε την δεύτερη θέση της ψηφοφορίας. Όμως το 2010 το VK πήρε το βραβείο «Trend-2009» από την Google.

Μια από τις σημαντικότερες διαφορές του VK από τα υπόλοιπα κοινωνικά δίκτυα ως το 2008 ήταν η παντελής έλλειψη διαφημίσεων. Ο Ντούροβ δεν ήταν ποτέ απόλυτος στις δηλώσεις του ως προς αυτό το θέμα, ώσπου τον Ιούλιο του 2008 υπέγραψε το συμβόλαιο με την «Media Plus». Στην συνέχεια η ηλεκτρονική διεύθυνση μετατράπηκε σε vk.com και βγήκε στην παγκόσμια αγορά, ενώ το 2010 έγινε μετάφραση σε δεκάδες γλώσσες.

Το 2008 ο Πάβελ Ντούροβ δημιούργησε το API δίνοντας την δυνατότητα σε προγραμματιστές και μη να χρησιμοποιούν τα δεδομένα με εύκολο τρόπο. Μετά το 2010 η δομή της σελίδας άλλαξε, οι χρήστες μπορούσαν όχι μόνο να βρίσκουν τους γνωστούς τους και να επικοινωνούν μαζί τους, αλλά να αναρτούν δημοσιεύσεις,

φωτογραφίες και μουσική. Με λίγα λόγια ένας απλός κοινωνικός ιστότοπος μετατράπηκε σε διαδικτυακό κέντρο ψυχαγωγίας και ενημέρωσης.

Το 2016 το VK ταυτίστηκε πλήρως με το Facebook, ενώ οι τελευταίες ανανεώσεις ανέβασαν το πιο γνωστό ρωσικό κοινωνικό δίκτυο στην δέκατη πέμπτη θέση παγκόσμιας επισκεψιμότητας. Σύμφωνα με τις αναλύσεις της στατιστικής εταιρείας Alexa [1]:

Σχήμα 0.1 Γραφική απεικόνιση επισκεψιμότητας του «Vkontakte»

Η αξιολόγηση των ιστότοπων πραγματοποιείται με συγκεκριμένα κριτήρια, όπως ο μέσος όρος των επισκέψεων ανά εικοσιτετράωρο, καθώς και το σύνολο των επισκέψεων των τελευταίων 90 ημερών. Παράλληλα οι αναλυτές αξιολογούν πάνω από 30 εκατομμύρια ιστότοπους παγκοσμίως και αξίζει να τονιστεί πως το VK έχει ξεπεράσει τα πιο δημοφιλή κοινωνικά δίκτυα όπως το Instagram, το οποίο βρίσκεται στην δέκατη έβδομη θέση, αλλά και το LinkedIn στην εικοστή όγδοη θέση στην παγκόσμια κατάταξη. Οι πρώτες τρεις θέσεις εδώ και πολύ καιρό ανήκουν στους:

- Google
- Youtube
- Facebook

Όσον αφορά τις προτιμήσεις του ρωσόφωνου πληθυσμού, την πρωτιά αναμφισβήτητα καταλαμβάνει το VKontakte. Σύμφωνα με την Alexa ακολουθούν:

- Ρωσική εκδοχή της Google

- Γιάντεξ (Яндекс)
- Youtube
- Mail.ru

Παρόλα αυτά, το δεύτερο δημοφιλέστερο κοινωνικό δίκτυο της Ρωσίας «Συμμαθητές» κατέχει μονάχα την έκτη θέση στην κατάταξη μεταξύ των Ρώσων χρηστών.

2.2 Χρήση του API

Το Application Programming Interface (API) είναι η διασύνδεση που βοηθάει στην αλληλεπίδραση του προγραμματιστή με οποιοδήποτε περιβάλλον που εμπεριέχει τα δεδομένα που τον ενδιαφέρουν. Αυτή η διεπαφή διευκολύνει σε μεγάλο βαθμό το χτίσιμο του κώδικα την στιγμή που προσφέρει έτοιμες κλάσεις και συναρτήσεις για την μελέτη των δεδομένων.

Συγκεκριμένα το API VKontakte είναι η διασύνδεση που επιτρέπει την λήψη πληροφοριών από την βάση δεδομένων του **vk.com** με την βοήθεια συγκεκριμένων ερωτημάτων που απευθύνονται στον server του. Δεν χρειάζεται λεπτομερής γνώση για την δομή της βάσης, των πινάκων ή των πεδίων που την αποτελούν, αρκεί που το API-αίτημα γνωρίζει όλα τα παραπάνω. Είναι άκρως σημαντικό τα ερωτήματα να είναι γραμμένα συντακτικά σωστά, ακολουθώντας αυστηρά τις οδηγίες.

2.2.1 Εξουσιοδότηση του χρήστη

Όπως όλα τα κοινωνικά δίκτυα, έτσι και το VK, προσφέρει την δυνατότητα στους χρήστες να ρυθμίζουν το απόρρητο των πληροφοριών και δημοσιεύσεων που αναρτούν, αλλά και να δημιουργούν την προσωπική τους μαύρη λίστα. Οι πληροφορίες που ορίζονται ως δημόσιες είναι ορατές για όλους, υπάρχουν και αυτές που προορίζονται μόνο για τους φίλους ή εκείνες που είναι προσβάσιμες αποκλειστικά και μόνο από τον ίδιο τον χρήστη. Με τις ίδιες αρχές λειτουργεί το API, με άλλα λόγια τα API-αιτήματα δεν μπορούν να φέρουν δεδομένα που είναι ιδιωτικά, για κάτι τέτοιο απαιτείται εξουσιοδότηση. Δηλαδή ο server πρέπει να γνωρίζει ποιος κάνει το αίτημα και σε τι πληροφορίες έχει πρόσβαση ο συγκεκριμένος χρήστης στη βάση του VK.

Αναλόγως τι είδους δεδομένα χρειάζεται ο χρήστης, επιλέγει την κατάλληλη μέθοδο. Πριν προχωρήσει κανείς στην εκμάθηση του API, απαιτείται δημιουργία μιας καινούργιας εφαρμογής όπου η διαδικασία είναι πολύ απλή και δεν απαιτεί πάνω από ένα λεπτό. Στην συνέχεια δίνεται το αναγνωριστικό του χρήστη, δηλαδή το ID της εφαρμογής που θα ζητείται μετέπειτα στις μεθόδους ως **app_id** για λήψη δεδομένων. Για την ακρίβεια καλείται η μέθοδος vk.API στην οποία δηλώνονται τα προσωπικά στοιχεία, το app_id, το user_login και το user_password. Ύστερα από αυτό, καλείται η επιθυμητή μέθοδος με μια συγκεκριμένη ονομασία (string), με τα πεδία που απαιτούνται (objects) και την συνάρτηση που θα επιστρέφει την απάντηση στα ερωτήματα.

2.2.2 VK-API για την Python

Για την υλοποίηση οποιουδήποτε προβλήματος με την βοήθεια του API χρειάζεται μια γλώσσα προγραμματισμού [16]. Τελευταία προτιμάται πολύ η Python, η οποία είναι σχετικά πιο ευέλικτη από τις άλλες γλώσσες υψηλού επιπέδου. Αφού λοιπόν η επιλογή της γλώσσας είναι σχεδόν αυτονόητη, το πρώτο βήμα είναι η εγκατάσταση της βιβλιοθήκης με την εντολή:

```
pip install vk
```

Συνεχίζοντας με την εξουσιοδότηση, κάποια δεδομένα είναι προσβάσιμα χωρίς την συμπλήρωση των προσωπικών στοιχείων, για παράδειγμα:

```
import vk
session = vk.Session()
vk_api = vk.API(session)
vk_api.users.get(user_id=1)
```

Κατά αυτόν τον τρόπο με την βοήθεια της μεθόδου **users.get** ανακαλείται το επίθετο, το όνομα και το id του χρήστη με το αναγνωριστικό αριθμό 1. Για περεταίρω πληροφορίες χρειάζεται επιπρόσθετη συμπλήρωση πεδίων, όπως [21]:

```
Vk_api.users.get(user_id=1, fields='city, country')
```

Δηλαδή στην προκειμένη περίπτωση, πέρα από το ονοματεπώνυμο του χρήστη με id=1, εμφανίζεται από ποια πόλη και χώρα είναι. Όμως χωρίς την εξουσιοδότηση δεν είναι εφικτή η μέγιστη δυνατή λειτουργικότητα του API. Έτσι υπάρχουν δύο τρόποι,

που είναι η είσοδος με το login και κωδικό του χρήστη ή με την χρήση token. Ο δεύτερος τρόπος ακολουθεί το παραπάνω παράδειγμα με ένα μικρό επιπρόσθετο στοιχείο:

```
session = vk.Session(access_token='token')
```

Ο άλλος τρόπος προϋποθέτει συμπλήρωση του login, κωδικού του χρήστη και κωδικού της εφαρμογής:

```
session = vk.AuthSession('app_id', 'user_login', 'user_password')

vk_api = vk.API(session)
```

Πέρα από την εξουσιοδότηση χρειάζεται διευκρίνιση των ζητούμενων πληροφοριών. Αυτό επιτυγχάνεται με την βοήθεια των κατάλληλων μεθόδων με τα αντίστοιχα συμπληρωμένα πεδία όπως φαίνεται παρακάτω:

```
session = vk.AuthSession('app_id', 'user_login', 'user_password')

vk_api = vk.API(session)

vk_api.wall.get(oauth='', owner_id='')
```

2.3 Περιγραφή των δεδομένων που έχουν συλλεγεί

Για την επίλυση οποιουδήποτε προβλήματος απαιτούνται τα κατάλληλα δεδομένα που θα βοηθήσουν στην λήψη αποφάσεων. Στα πλαίσια της συγκεκριμένης διπλωματικής μελέτης ερευνήθηκαν τα θέματα που ενδιαφέρουν και απασχολούν τους Ρώσους πολίτες σε σχέση με την Ελλάδα, έτσι ώστε ο τουριστικός κλάδος να στραφεί στοχευμένα στην βελτιστοποίηση των υπηρεσιών.

Για την αρχή συλλέχθηκαν τα δεδομένα από το κοινωνικό δίκτυο VK με την βοήθεια του API στην python με την διαδικασία που περιγράφηκε προηγουμένως. Χρειάστηκαν όλες οι σελίδες, η θεματολογία των οποίων συσχετιζόταν με την Ελλάδα (τουρισμός, πολιτική, κοινωνία και γαστρονομία), οι ομάδες με μικρό αριθμό μελών δεν λήφθηκαν υπόψιν. Συνολικά συγκεντρώθηκαν 49.671 δημοσιεύσεις από 12 σελίδες που εμπεριείχαν 201.401 μέλη.

```
In [64]: len(open('new_final_final.csv').readlines())
Out[64]: 49671
```

Σχήμα 0.2 Αριθμός δημοσιεύσεων

Για την συλλογή αυτών των δημοσιεύσεων χρειάστηκαν τα id των επιλεγμένων σελίδων που εμφανίζονται με την μέθοδο groups.search. Για παράδειγμα το id της ομάδας «Греция / Greece / Ελλάδα» τυπώνεται ως εξής:

```
In [66]: group = api.groups.search(q='Греция / Greece / Ελλάδα')

In [67]: print group[1]['gid']
5628094
```

Σχήμα 0.3 API μέθοδος για την εύρεση του id της ομάδας «Vkontakte»

Στην συνέχεια μαζεύονται οι δημοσιεύσεις και τα σχόλια αυτών, αν υπάρχουν, από τα id των ζητούμενων σελίδων. Η συλλογή τους γίνεται με την βοήθεια δύο μεθόδων: wall.get και wall.getComments αντίστοιχα. Στην πρώτη περίπτωση η λήψη του κειμένου προϋποθέτει τον αριθμό εξουσιοδότησης και το id της σελίδας στην οποία αυτό αναρτήθηκε.

```
In [60]: posts = api.wall.get(oauth='5369754', owner_id='-34779244',
 count=10)

In [61]: print posts[3]['text']
Фестиваль уличной еды пройдёт в Салониках !<br> <br>Двухдневный Фест
иваль уличной еды Thessaloniki Street Food Festival пройдет в Салони
ках 29 и 30 апреля. Знаменитая салоникская выпечка - «бугаца», «кулу
ри», а также другие известные блюда северной столицы еще раз поразят
 своим разнообразием - ведь Салоники, как говорят его жители, самый
мультикультурный город Греции. Фестиваль еды состоит из двух отделен
ий - Street Food Area, где представят рестораны, таверны, кафе, и St
reet Market Area, где будет представлена продукция - вино, мед, масл
о. Адрес: городская мэрия (новое здание) Салоник <br><br>начало в 10
:00 до 22:00
```

Σχήμα 0.4 Εμφάνιση της δημοσίευσης από την σελίδα του «Vkontakte»

Στην δεύτερη περίπτωση δεν είναι αναγκαία η εξουσιοδότηση του προγραμματιστή, όμως για την λήψη των σχολίων χρειάζεται το id της δημοσίευσης στην οποία αναφέρεται το σχόλιο αυτό.

```
In [51]: comments = api.wall.getComments(owner_id='-3477924
4', post_id=posts[3]['id'])

In [52]: print comments[1]['text']
Обучение греческому языку с нуля. <br>Не упустите ТАКУЮ зам
ечательную возможность!Быстрее берите телефон в руки и запи
сывайтесь!♫. Занятия провожу как для детей, так и для взро
слых. Атмосфера- ☼Помогу Вам выучить греческий быстро и ка
чественно с помощью аудио-,видео материалов,большой базы ме
тодической и учебной литературы, поставлю правильное произн
ошение. По Очень низкой цене. Пишите в ЛС. Занятия по Skype
```

Σχήμα 0.5 Εμφάνιση του σχολίου της ζητούμενης δημοσίευσης

Εν τέλει, τα κείμενα αποθηκεύονται σε ένα αρχείο .csv το οποίο στην συνέχεια θα αποτελέσει την πηγή των δεδομένων προς ανάλυση. Η διαδικασία της μετατροπής των δημοσιεύσεων σε κατάλληλη μαθηματική μορφή θα αναλυθεί στο επόμενο κεφάλαιο.

3. Επεξεργασία φυσικής γλώσσας (ΕΦΓ)

Οι πιο διαδεδομένες μέθοδοι στην Ανάλυση Δεδομένων μπορούν να επεξεργάζονται μόνο μαθηματικά μεγέθη, επομένως όταν απαιτείται ανάλυση των κειμένων, τίθεται πρόβλημα μετατροπής των λέξεων σε κατάλληλη αναπαράσταση [14]. Την λύση αυτού του προβλήματος έρχεται να δώσει μια νέα τεχνολογία, η οποία δημιουργήθηκε από τον Tomas Mikolov στην έδρα της Google. Η βασική ιδέα βασίζεται στην κωδικοποίηση των λέξεων, αλλά και ολόκληρων παραγράφων ως μαθηματική αναπαράσταση πάνω σε διανυσματικούς χώρους, με άλλα λόγια κάθε λέξη αντιστοιχίζεται σε μια μοναδική προβολή ή διάνυσμα. Η μέθοδος αυτή είναι χρήσιμη σε προβλήματα όπως [19]:

- Θεματική ομαδοποίηση
- Συναισθηματική Ανάλυση
- Φιλτράρισμα της ανεπιθύμητης αλληλογραφίας
- Σύσταση διαφημίσεων
- Μηχανές αναζήτησης
- Ομαδοποίηση Εγγράφων
- Ταξινόμηση Εγγράφων
- Μηχανική μετάφραση

Για την επίτευξη των παραπάνω εφαρμογών χρησιμοποιούνται αλγόριθμοι ταξινόμησης ή ομαδοποίησης. Καθώς τα δεδομένα πρέπει να αναπαρίστανται ως αριθμοί, οι πιο συνηθισμένοι τρόποι μετατροπής κειμένων σε μαθηματικές αναπαραστάσεις είναι η bag-of-words ή η bag-of-n-grams.

Τα μειονεκτήματα της bag-of-words είναι αρκετά και σημαντικά, το κυριότερο είναι η απουσία του συντακτικού. Την στιγμή που δεν λαμβάνεται υπόψη η σειρά των λέξεων, δυο προτάσεις με διαφορετική σημασία αποτελούν το ίδιο αποτέλεσμα, αρκεί να περιέχουν τις ίδιες λέξεις. Από την άλλη το bag-of-n-grams περιέχει την έννοια του συντακτικού, όμως δημιουργεί τεράστιες διαστάσεις. Κανένας όμως από τους προαναφερόμενους τρόπους δεν αντιλαμβάνεται την σημασιολογία των λέξεων. Για παράδειγμα, οι λέξεις «όμορφος», «ωραίος» και «θάλασσα» έχουν ίσες αποστάσεις μεταξύ τους, ενώ σημασιολογικά «όμορφος» θα έπρεπε να είναι πιο κοντά στο «ωραίος» από ότι στην «θάλασσα».

Η επεξεργασία της φυσικής γλώσσας είναι ένας διεπιστημονικός κλάδος της μηχανικής μάθησης. Ένας από τους παλαιότερους τρόπους μετατροπής των λέξεων σε διανύσματα είναι η λεγόμενη one-hot-κωδικοποίηση ή αφελής (dummy) κωδικοποίηση. Για την κάθε μια λέξη δημιουργείται ένας δυαδικός συνδυασμός που αποτελείται από **N** στοιχεία, ο οποίος δηλώνει τον συνολικό αριθμό μοναδικών λέξεων που έχουν εμφανιστεί στο σύνολο των κειμένων (vocabulary size). Για παράδειγμα :

Ρώμη = [1, 0, 0, 0, 0, 0, … , 0]

Παρίσι = [0, 1, 0, 0, 0, 0, …, 0]

Ιταλία = [0, 0, 1, 0, 0, 0, …, 0]

Γαλλία = [0, 0, 0, 1, 0, 0, …, 0]

Σχήμα 3.1 Απεικόνιση της one-hot-κωδικοποίησης

Το μειονέκτημα της παραπάνω μεθόδου είναι το τεράστιο μέγεθος του λεξικού, που προκαλεί με την σειρά του πολλά προβλήματα στην ανάλυση. Η κατανεμημένη αναπαράσταση έρχεται να λύσει αυτά τα θέματα ορίζοντας τις διαστάσεις των διανυσμάτων πολύ μικρότερες από το μέγεθος του λεξικού [13].

Ρώμη = [0.91, 0.83, 0.17, … , 0.41]

Παρίσι = [0.92, 0.82, 0.17, … , 0.98]

Ιταλία = [0.32, 0.77, 0.67, … , 0.42]

Γαλλία = [0.33, 0.78, 0.66, … , 0.97]

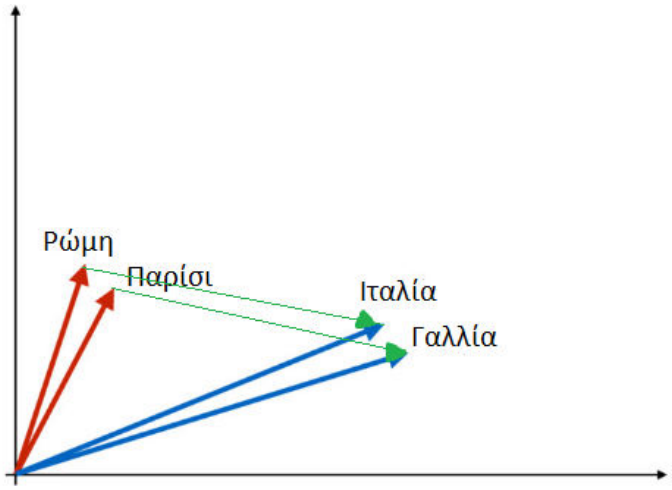
Με την κατανεμημένη αναπαράσταση φαίνεται πως η συγγένεια των λέξεων αποτυπώνεται στα διανύσματα :

Ρώμη = [0.91, 0.83, 0.17, … , 0.41]

Παρίσι = [0.92, 0.82, 0.17, … , 0.98]

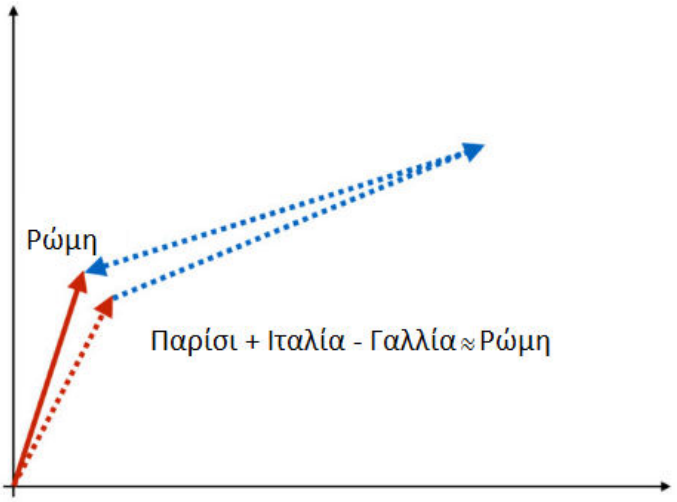
Ιταλία = [0.32, 0.77, 0.67, … , 0.42]

Γαλλία = [0.33, 0.78, 0.66, … , 0.97]



Σχήμα 3.2 Γραφική αναπαράσταση των διανυσμάτων των χωρών-πρωτευουσών

Παρατηρείται πως η σχέση του διανύσματος της Ρώμης ως προς το διάνυσμα της Ιταλίας είναι ίδια με εκείνη του διανύσματος του Παρισίου με την Γαλλία. Επίσης διακρίνεται ότι τα διανύσματα σχέσεων των πρωτευουσών ως προς τις χώρες που ανήκουν έχουν σχεδόν την ίδια κατεύθυνση, δηλαδή είναι σχεδόν παράλληλες. Αυτό δίνει την δυνατότητα να κάνουμε διανυσματικές πράξεις με τις λέξεις :



Σχήμα 3.3 Γραφική αναπαράσταση της πράξης των διανυσμάτων

Ένας άλλος τρόπος που χρησιμοποιείτο μέχρι πρότινος είναι η λεγόμενη στατιστική ΕΦΓ όπου ένα σύνολο εγγράφων **T** αποτελούνταν από έγγραφα **D** τα οποία με την σειρά τους αναπαρίστανται ως ακολουθίες λέξεων [7]:

$$D = \{x_1, \dots, x_n\}$$

Όλες οι μοναδικές λέξεις από το σύνολο εγγράφων **T** δημιουργούν λεξικό **V** και κάθε έγγραφο **D** επισημαίνεται με ετικέτα (label) **Y**, για παράδειγμα :

Είναι εκπληκτικός ο καιρός στην Χίο σήμερα ! → 1

@anna, Ήταν η χειρότερη ταινία που έχω δει → -1

Πάνω στα έγγραφα **D** ή στα ζεύγη (**D,Y**) χτίζονται μοντέλα πιθανολογικού χαρακτήρα όπου στην ουσία καταγράφεται η συχνότητα εμφάνισης της κάθε λέξης στο λεξικό. Όμως αυτό έχει αποτέλεσμα να χάνεται η σύνταξη, γεγονός που αποτελεί ένα σημαντικό μειονέκτημα.

Αν υποθέσουμε ότι έχουμε ένα λεξικό **V** :

(… , κακός = 140, φυτό = 141, … , άθλιος = 1299,…)

Ενώ οι λέξεις «κακός» και «άθλιος» είναι συνώνυμες, σύμφωνα με την παραπάνω αναπαράσταση δεν σχετίζονται μεταξύ τους. Πρακτικά από ένα σύνολο **T** είναι γνωστό ότι :

$$P(\text{“σκύλος ”} \mid \text{”άνθρωπος”} , \text{“φίλος”}) = 0{,}997$$

‘Όμως για λέξεις όπως “χάσκι”, ”παιδί” και ”κολλητός” δεν υπάρχουν τέτοιου είδους πληροφορίες. Στην προκειμένη περίπτωση δεν μπορούν να υπολογιστούν οι πιθανότητες όπως:

$$P(\text{“σκύλος”} \mid \text{”παιδί”} , \text{“φίλος”}) = ?$$

$$P(\text{“σκύλος”} \mid \text{”άνθρωπος”} , \text{“κολλητός”}) = ?$$

P(“χάσκι” | ”άνθρωπος” , “κολλητός”) = ?

Ο στόχος είναι να αποκτήσουν οι λέξεις τέτοια διανύσματα ώστε να ισχύει :

V(“άνθρωπος”)≈ V(“παιδί”)

V(“σκύλος”)≈ V(“χάσκι”)

V(“φίλος”)≈ V(“κολλητός”)

Τέτοιος συσχετισμός επιτυγχάνεται με την μετατροπή των λέξεων σε εμφυτεύσεις (word embeddings ή Word2vec).

Κατά συνέπεια θα ισχύει :

P(“σκύλος” | ”παιδί” , “φίλος”) ≈ P(“σκύλος” | ”άνθρωπος” , “φίλος”) ≈ 0,997

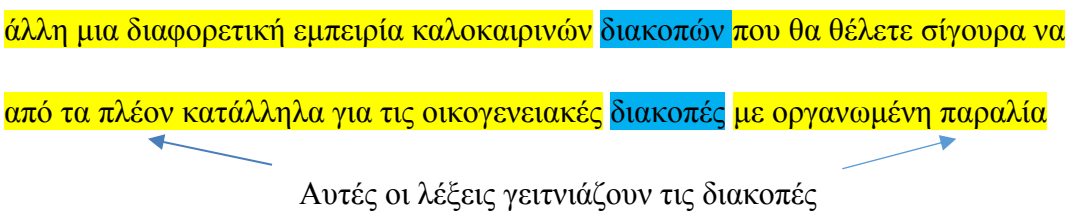
P(“χάσκι” | ”άνθρωπος” , “φίλος”) ≈ P(“σκύλος” | ”άνθρωπος” , “φίλος”) ≈ 0,997

P(“σκύλος” | ”παιδί” , “κολλητός”) ≈ P(“σκύλος” | ”άνθρωπος” , “φίλος”) ≈ 0,997

3.1 Η κατανεμημένη αναπαράσταση

Η μέθοδος Word2vec είναι ένα εργαλείο ή αλλιώς, ένας συνδυασμός αλγορίθμων, ο στόχος της οποίας είναι να αντιστοιχηθούν οι λέξεις σε τέτοια διανύσματα, ώστε οι σχέσεις μεταξύ των διανυσμάτων αυτών να αντιπροσωπεύουν τις σχέσεις μεταξύ των λέξεων στην φυσική γλώσσα. Με άλλα λόγια κυριαρχεί η ιδέα «Δείξε μου τις λέξεις σου να σου πω ποιος είσαι».

Για παράδειγμα :



Σχήμα 3.4 Αναπαράσταση των γειτονικών λέξεων

Στην ουσία η ταυτότητα της λέξης ορίζεται σύμφωνα με την θέση στην οποία βρίσκεται στην πρόταση και από τους γείτονές της. Συνήθως οι διακοπές ταυτίζονται

με τις λέξεις “καλοκαίρι”, “θάλασσα”, “παραλίες”, όπως φαίνεται στις παραπάνω δυο προτάσεις. Το Word2vec δεν ξέρει τι σημαίνουν οι διακοπές, το μόνο που γνωρίζει είναι τα συμφραζόμενα, δίπλα στα οποία εμφανίζονται. Πρακτικά η διαδικασία αυτή πραγματοποιείται σε τρία βήματα [11]:

- Επιλογή της κατάλληλης συνάρτησης
- Αρχικοποίηση τυχαίων διανυσμάτων
- Αλγόριθμος απότομης καθόδου (Gradient descent)

Ο στόχος του αλγορίθμου Word2vec είναι να υπολογίσει την πιθανότητα εμφάνισης μίας λέξης δίπλα σε μια άλλη. Οι λέξεις που βρίσκονται κοντά έχουν υψηλότερη πιθανότητα από άλλες δύο που δεν εμφανίζονται μαζί. Για παράδειγμα, έχοντας την παρακάτω πρόταση εξετάζεται η πιθανότητα εμφάνισης των γύρω λέξεων μαζί με το **over** [5]:

“The fox jumped **over** the lazy dog”

P(the | over)

P(fox | over)

P(jumped | over)

P(the | over)

P(lazy | over)

P(dog | over)

Ο στόχος είναι να βρεθούν οι πιθανότητες εμφάνισης των λέξεων όπως **the**, **fox** κλπ. στην ίδια πρόταση με το **over**. Αφού εντοπίστηκαν αυτές οι λέξεις κοντά στο **over**, η πιθανότητα τους αυξάνεται, ενώ η πιθανότητα των υπόλοιπων λέξεων ως προς το **over** μειώνεται. Φυσικά λέγοντας «πιθανότητα εμφάνισης της fox δίπλα στο over» εννοείται πιθανότητα εμφάνισης του διανύσματος fox όταν υπάρχει over στα συμφραζόμενα.

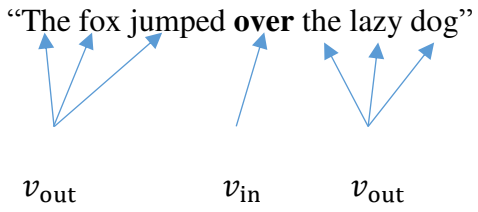
$$P(\text{fox} \mid \text{over})$$

$$P(v_{\text{fox}} \mid v_{\text{over}})$$

Σχήμα 3.5 Αναπαράσταση της πιθανότητας εμφάνισης της fox δίπλα στο over

Σε κάθε λέξη αντιστοιχούν δύο διανύσματα, εισόδου και εξόδου, και ανάλογα σε ποια θέση βρίσκεται η λέξη, επιλέγεται το κατάλληλο διάνυσμα :

$$P(v_{\text{out}} \mid v_{\text{in}})$$



Σχήμα 3.6 Αντιστοίχιση των διανυσμάτων εισόδου-εξόδου στις λέξεις

Ο υπολογισμός των πιθανοτήτων εμφάνισης με κεντρική την λέξη **over,** γίνεται με ανάθεση του διανύσματος εισόδου στην **over**, ενώ τα συμφραζόμενα αντιπροσωπεύονται από τα διανύσματα εξόδου. Παίρνοντας με την σειρά τις λέξεις που βρίσκονται γύρω από την κεντρική, αυξάνεται η πιθανότητά εμφάνισής τους λόγω της θέσης στην οποία βρίσκονται. Ύστερα μετακινείται το παράθυρο έχοντας την λέξη **the** αυτήν την φορά ως κεντρική και ο υπολογισμός θα γίνεται ακολουθώντας τα ίδια βήματα. Η ομοιότητα δύο λέξεων υπολογίζεται με την βοήθεια του εσωτερικού γινομένου των διανυσμάτων αναπαράστασής τους:

$$v_{\text{out}} * v_{\text{in}}$$

Άλλος τρόπος υπολογισμού της ομοιότητας είναι το συνημίτονο της γωνίας που σχηματίζεται μεταξύ τους:

$$\text{Ομοιότητα} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}||||\mathbf{B}||} = \frac{\sum_{i=1}^{n} \mathrm{A_i} \times \mathrm{B_i}}{\sqrt{\sum_{i=1}^{n} (\mathrm{A_i})^2} \times \sqrt{\sum_{i=1}^{n} (\mathrm{B_i})^2}}$$

Το αποτέλεσμα της παραπάνω σχέσης προφανώς κυμαίνεται στο εύρος τιμών [-1, 1]. Για να υπολογιστεί η πιθανότητα με εύρος τιμών [0, 1] χρησιμοποιείται η συνάρτηση softmax :

$$\text{softmax} = \frac{\exp(\mathrm{v_{in}} \cdot \mathrm{v_{out}})}{\sum_{\mathrm{k}\in\mathrm{v}} \exp(\mathrm{v_{in}} \cdot \mathrm{v_k})} = \mathrm{P}(v_{\mathrm{out}} \mid v_{\mathrm{in}})$$

Με αυτόν τον τρόπο υπολογίζεται η ομοιότητα μεταξύ των λέξεων **fox** και **over**. Ο αριθμητής αποτελείται από την εκθετική συνάρτηση αυτών των δύο διανυσμάτων και ο παρονομαστής είναι η εκθετική συνάρτηση του αθροίσματος όλων των συνδυασμών του **over** με όλες τις λέξεις που υπάρχουν στο λεξικό. Επομένως αν το **fox** και **over** βρίσκονται συχνά στην ίδια πρόταση πρέπει διανυσματικά επίσης να είναι κοντά, δηλαδή το εσωτερικό τους γινόμενο να είναι μεγάλο. Το εσωτερικό γινόμενο είναι στην πραγματικότητα η γωνία που σχηματίζεται ανάμεσα στα δύο σημεία, όταν αυτά βρίσκονται κοντά το ένα στο άλλο, τότε το εσωτερικό γινόμενο είναι μεγάλο, ενώ όταν η απόσταση μεταξύ των σημείων είναι μεγάλη, το γινόμενο είναι μικρό. Είναι πολύ ενδιαφέρον όταν τα διανύσματα είναι αντίθετα, το εσωτερικό τους γινόμενο ισούται με -1, έτσι μπορούν να εντοπιστούν οι αντίθετες λέξεις.

Ενώ το Word2vec μπορεί και εντοπίζει τα συνώνυμα και είναι δυνατές οι διανυσματικές πράξεις με τις λέξεις, θα μπορούσε κανείς να σκεφτεί πως είναι πιθανόν να προκύψει το αντώνυμο της λέξης, αν αυτή πολλαπλασιαστεί με -1. Δεν ισχύει όμως κάτι τέτοιο, καθώς η κατανεμημένη αναπαράσταση υπολογίζει την σημασιολογική ομοιότητα των λέξεων, χωρίς να κατανοεί αν οι δύο λέξεις είναι αντώνυμα. Για παράδειγμα :

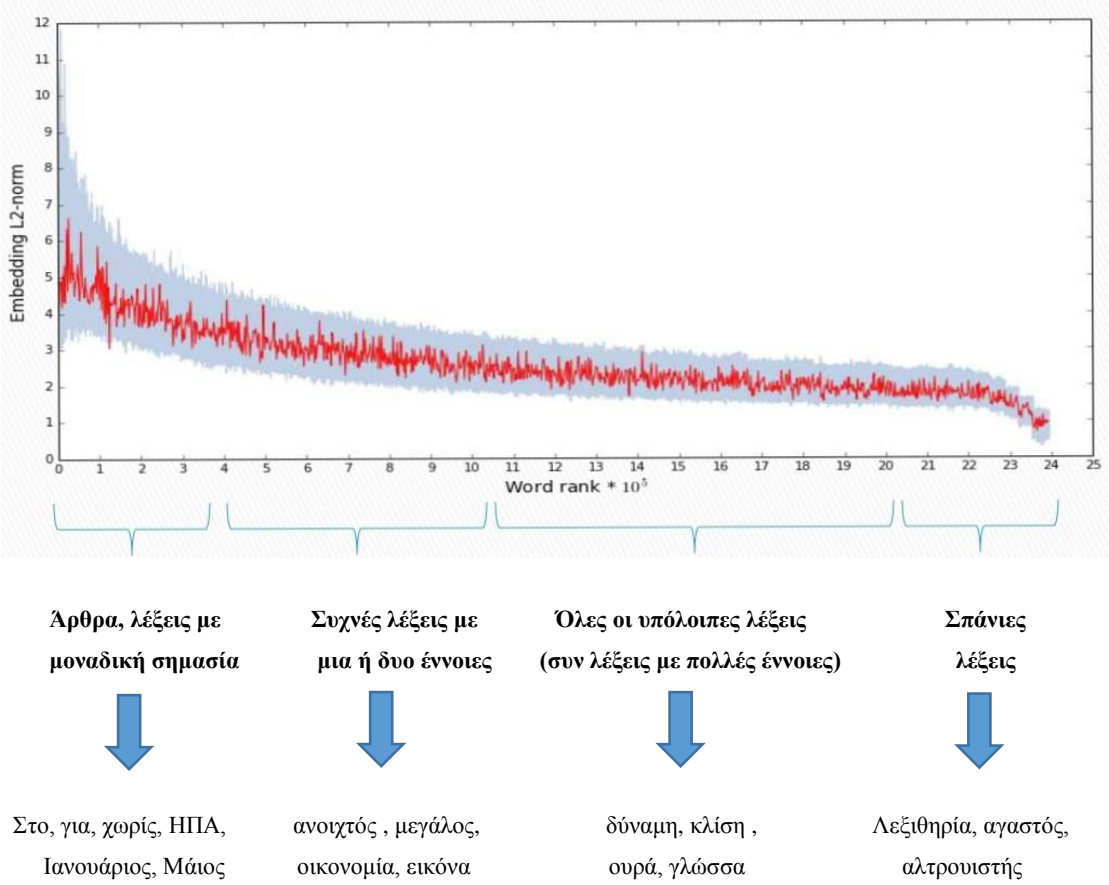
$$\text{"good"} \neq (-1) * \text{" bad"}$$

Αντίστοιχα τα διανύσματα εισόδου και εξόδου δεν μοιάζουν μεταξύ τους και προφανώς παραλείποντας μια από τις δυο κατηγορίες διανυσμάτων, μπορεί να χειροτερεύσει το μοντέλο.

$$\text{"good"}_{\mathrm{in}} \neq \text{"good"}_{\mathrm{out}}$$

Μια άλλη παρατήρηση για το μήκος (norm) των διανυσμάτων, προκύπτει από το παρακάτω διάγραμμα, όπου οι λέξεις έχουν ταξινομηθεί ως προς την συχνότητα

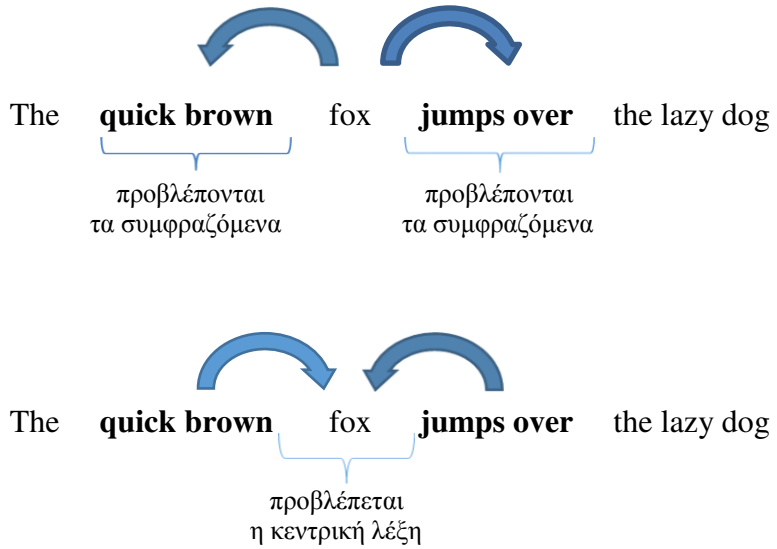
εμφάνισής τους, όπως φαίνεται στον άξονα **x** και στον άξονα **y** που είναι τα αντίστοιχα μήκη τους.



Σχήμα 3.7 Γραφική αναπαράσταση του καταμερισμού εμφάνισης των λέξεων
Πηγή: https://www.slideshare.net/DenisDus1/word2vec-part-1-61489929

Φαίνεται πως στις λέξεις που εμφανίζονται πιο συχνά ή έχουν μοναδική σημασία αντιστοιχούν τα πιο μεγάλα διανύσματα. Ενδιαφέρον αποτελούν και οι σπάνιες λέξεις, διότι η αρχικοποίηση των διανυσμάτων γίνεται με τυχαίο τρόπο και σε περίπτωση που δεν ξαναεμφανιστούν οι λέξεις στο κείμενο, θα εξακολουθούν να είναι σχεδόν τυχαίες.

Πιο αναλυτικά, το κείμενο διασπάται σε επιμέρους υποπροτάσεις, η πρόβλεψη γίνεται είτε για τις γύρω λέξεις γνωρίζοντας την κεντρική (Skip-gram), ή για την κεντρική λέξη δεδομένου μερικών λέξεων που βρίσκονται αριστερά και δεξιά (Continuous Bag of Words).



Σχήμα 3.8 Απεικόνιση του παραθύρου word2vec
Πηγή: https://www.slideshare.net/DenisDus1/word2vec-part-1-61489929

Η έννοια της μετατροπής των λέξεων σε διανύσματα δεν είναι εντελώς καινούργια, όμως η σπουδαιότητα της μεθόδου Word2vec και Doc2vec του Tomas Mikolov μπορεί να διαπιστωθεί από 7313 αναφορές σε αυτήν σε διαφορετικά άρθρα. Το ερώτημα είναι πως επιτυγχάνεται το επιθυμητό αποτέλεσμα. Όπως προαναφέρθηκε χρειάζεται ένα μεγάλο σύνολο εγγράφων **T** (για παράδειγμα η Google έχει χρησιμοποιήσει όλα τα δεδομένα της Wikipedia), πάνω στο οποίο χτίζονται δείκτες για όλες τις λέξεις π.χ. της αγγλικής γλώσσας στο λεξικό **V,** τέτοιοι ώστε να μεγιστοποιούν την πιθανότητα εμφάνισης τους στο συνολικό **T**.

Δηλαδή : $\theta^* = \underset{\theta}{\operatorname{argmax}}\, P(T,\theta)$

παράμετροι του μοντέλου (δείκτες ή διανύσματα)

πιθανότητα εμφάνισης συνόλου T δεδομένου δεικτών θ

Στην συνέχεια αναλύεται η παραπάνω σχέση με σειρά από ενδεχόμενα :

$$P(\theta \mid T) = \prod_{D \in T} P(\theta \mid D) = \prod_{D \in T} \prod_{w \in D} (P(\theta \mid w, c(w)) =$$

1. Τα έγγραφα στο σύνολο T είναι ανεξάρτητα μεταξύ τους
2. Κάθε λέξη στο έγγραφο εξαρτάται μόνο από τα συμφραζόμενα

$$= \prod_{D \in T} \prod_{w \in D} \prod_{c \in C(w)} (P(\theta \mid w, c) = \prod_{D \in T} \prod_{w \in D} \prod_{c \in C(w)} \frac{\exp(v_w^T v_c)}{\Sigma_{c' \in V} \exp(v_w^T v_{c'})} \rightarrow \max_{\{v_w, v_c\}}$$

3. Οι λέξεις από τα συμφραζόμενα δεν συνδέονται μεταξύ τους

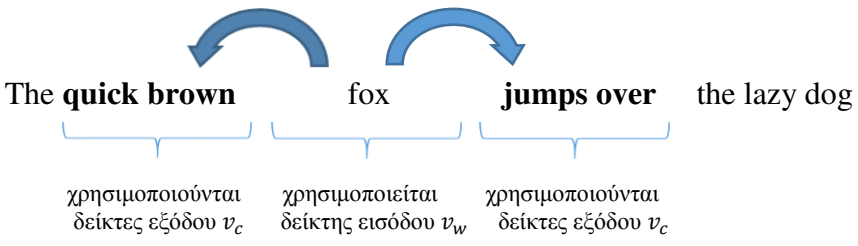
4. Πιθανότητα εμφάνισης των (w,c) δημιουργείται με την βοήθεια της συνάρτησης Softmax

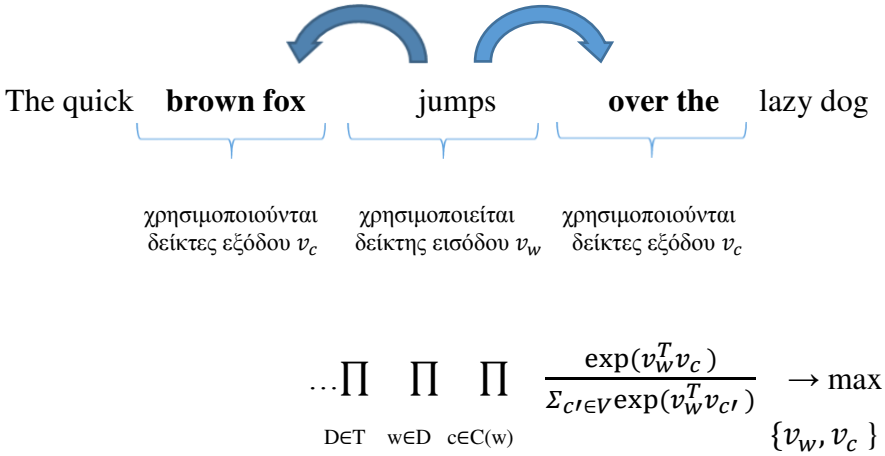
Έχοντας :

1. Παράμετροι του μοντέλου $\theta = \{ W, W' \}$
2. W – πίνακας δεικτών εισόδου
3. W΄– πίνακας δεικτών εξόδου
4. v_w – δείκτης εισόδου της κεντρικής λέξης **w**
5. v_c – δείκτης εξόδου της λέξης **c** των συμφραζομένων

} Κάθε λέξη διαθέτει δύο διανύσματα

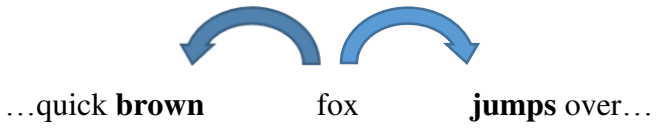
Με την βοήθεια του προηγούμενου παραδείγματος διαπιστώνεται η ιδιαιτερότητα της μεθόδου Word2vec η οποία χρησιμοποιεί δύο διαφορετικά διανύσματα για την κάθε λέξη. Ανάλογα με την περίπτωση που εξετάζεται και σε ποια θέση βρίσκεται η λέξη, εφαρμόζεται το κατάλληλο διάνυσμα. Με την βοήθεια του κινούμενου παραθύρου πραγματοποιείται η σάρωση του κειμένου με τελικό στόχο την πρόβλεψη των συμφραζομένων, δεδομένου της κεντρικής λέξης. Δηλαδή, το παράθυρο μετακινείται έχοντας πρώτα ως κεντρική την λέξη fox στην οποία αντιστοιχεί ο δείκτης εισόδου και ως συμφραζομένων τις δυο λέξεις αριστερά και δεξιά οι οποίες αποκτούν τους δείκτες εξόδου. Έπειτα γίνεται ξανά η μετακίνηση με κεντρική την jumps που θα εμφανίζεται αυτήν την φορά με τον δείκτη εισόδου, ενώ η fox με τον εξόδου.





Σχήμα 3.9 Απεικόνιση της μεταφοράς του παραθύρου word2vec
Πηγή: https://www.slideshare.net/DenisDus1/word2vec-part-1-61489929

Κατά αυτόν τον τρόπο υπολογίζεται η πιθανότητα εμφάνισης μιας συγκεκριμένης λέξης δίπλα σε μια άλλη δεδομένου του λεξικού **V**. Για παράδειγμα αν υπάρχει ένα μοναδικό έγγραφο που αποτελείται από 5 λέξεις, καθώς και το λεξικό να αποτελείται από 5 λέξεις, οι υπολογισμοί θα γίνουν με τον εξής τρόπο :



Σχήμα 3.10 Απεικόνιση του θεωρητικού λεξικού

1. Λεξικό V = { quick brown fox jumps over }
2. $v_w^T v_c$ - πραγματικός αριθμός που αντιστοιχεί στο διάνυσμα

3. Έστω :

$$v_{fox}^T v_{brown} = 17.7 \quad \frac{\exp(17.7)}{\exp(17.7)+\exp(15.2)+\exp(11.3)+\exp(7.9)} \approx 0.9227$$

$$v_{fox}^T v_{jumps} = 15.2 \quad \frac{\exp(15.2)}{\exp(17.7)+\exp(15.2)+\exp(11.3)+\exp(7.9)} \approx 0.0757$$

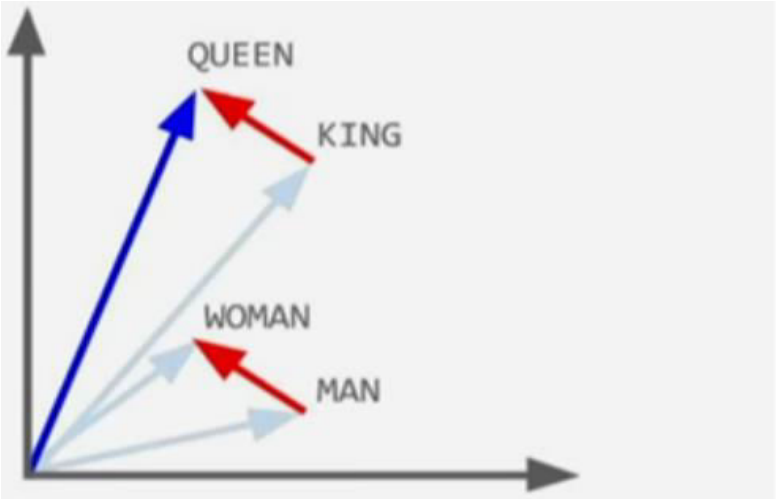
$$v_{fox}^T v_{quick} = 11.3 \quad \frac{\exp(11.3)}{\exp(17.7)+\exp(15.2)+\exp(11.3)+\exp(7.9)} \approx 0.0015$$

$$v_{fox}^T v_{over} = 7.9 \quad \frac{\exp(7.9)}{\exp(17.7)+\exp(15.2)+\exp(11.3)+\exp(7.9)} \approx 0.0001$$

$$\Big\} = 1$$

$$\ldots \prod_{D \in T} \prod_{w \in D} \prod_{c \in C(w)} \frac{\exp(v_w^T v_c)}{\Sigma_{c' \in V} \exp(v_w^T v_{c'})} \rightarrow \max_{\{v_w, v_c\}}$$

Το κλασικό παράδειγμα του αλγορίθμου κατανεμημένης αναπαράστασης αποτελεί το διάνυσμα, το οποίο προκύπτει από την αφαίρεση του διανύσματος της λέξης «άνδρας» (man) από το διάνυσμα της λέξης «γυναίκας» (woman) :



Σχήμα 3.11 Αναπαράσταση του διανύσματος της θηλυκότητας

Άρα, αν στο διάνυσμα της λέξης «βασιλιάς» (king) προστεθεί αυτό το διάνυσμα, δηλαδή **KING – MAN + WOMAN**, θα προκύψει το διάνυσμα της «βασίλισσας» (queen). Με αυτόν τον τρόπο αναδεικνύεται η άμεση σύνδεση της φυσικής γλώσσας

με την μαθηματική αναπαράστασή της, προσφέροντας την δυνατότητα πρόσθεσης ή αφαίρεσης λέξεων σαν να είναι απλοί αριθμοί :

Russian + river = Moscow

Vietnam + capital = Hanoi

Μια πολύ σημαντική παρατήρηση που αναφέραμε προηγουμένως αποτελεί το γεγονός ότι το Word2vec εντοπίζει τα συνώνυμα των λέξεων, ενώ δεν ισχύει το ίδιο για τα αντώνυμα. Αυτό αφορά κυρίως στα αγγλικά, διότι εκεί οι λέξεις δεν διαφέρουν ιδιαίτερα όταν αλλάζει ο χρόνος ή το πρόσωπο. Ως προς τα ρωσικά είναι δύσκολη η μορφοποίηση, γιατί η ρίζα της λέξης αλλάζει σημαντικά στην αλλαγή του γένους, χρόνου, αριθμού ή προσώπου, κάνοντας το μοντέλο λιγότερο ακριβές και λιγότερο ευέλικτο.

3.1.1 Αρχιτεκτονική του Word2vec

Η κατανεμημένη αναπαράσταση λειτουργεί με δύο βασικές αρχιτεκτονικές που είναι Continuous Back of Words (CBOW) και Skip-gram που είδαμε προηγουμένως. Πολλοί θεωρούν πως το Word2vec είναι αλγόριθμος βαθιάς μάθησης (Deep learning), κάτι που δεν ισχύει μιας και χρησιμοποιείται ένα ρηχό (shallow) νευρωνικό δίκτυο πρόσω-τροφοδότησης (ενός επιπέδου). Η διαδικασία που ακολουθείται αποτελείται από τα εξής βήματα :

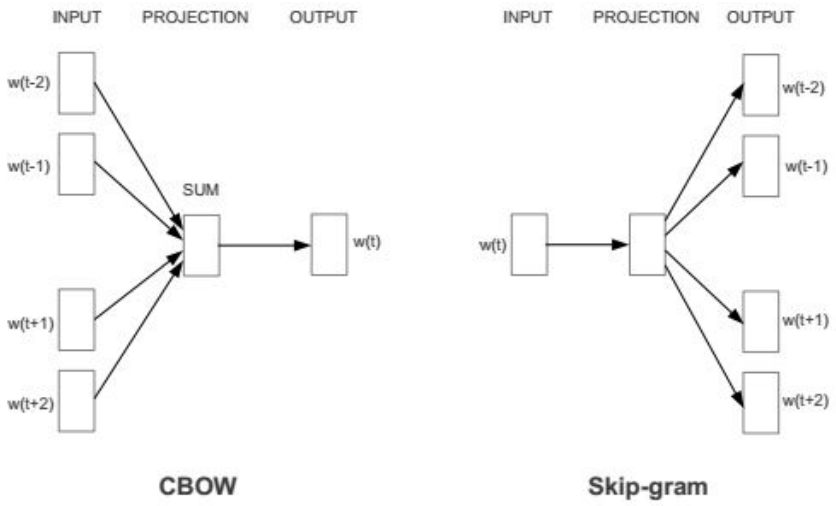
1. Διαβάζεται ένα δυαδικό δέντρο Huffman κείμενο και υπολογίζεται η συχνότητα εμφάνισης της κάθε λέξης, δηλαδή πόσες φορές εμφανίζεται η κάθε λέξη στο κείμενο
2. Δημιουργείται πίνακας στον οποίο ταξινομούνται οι λέξεις σύμφωνα με την συχνότητα εμφάνισης τους. Οι λέξεις που εμφανίζονται μόνο μια φορά διαγράφονται
3. Χτίζεται ένα διαδικό δέντρο Χάφμαν (Huffman Binary tree) που συχνά χρησιμοποιείται στους αλγορίθμους κωδικοποίησης του λεξικού
4. Έπειτα το κείμενο χωρίζεται σε υποπροτάσεις που μπορεί να είναι η ίδια η πρόταση ή ολόκληρο κομμάτι κειμένου. Ύστερα γίνεται δειγματοληψία κατά την οποία εντοπίζονται οι πιο συχνές λέξεις, κάτι που επιταχύνει την

διαδικασία εκπαίδευσης του αλγορίθμου και καλυτερεύει την ακρίβεια του τελικού μοντέλου

5. Υπάρχει το λεγόμενο παράθυρο το οποίο μετακινείται πάνω στις υποπροτάσεις. Το παράθυρο αυτό είναι μια παράμετρος που ορίζεται στον κώδικα, στην ουσία είναι η μέγιστη απόσταση μεταξύ της λέξης που εξετάζεται και των συμφραζομένων. Για παράδειγμα, αν το παράθυρο ισούται με τέσσερα, στην πρόταση «Είναι υπέροχο το καλοκαίρι στην Ελλάδα», η ανάλυση του κειμένου (επιλογή αλγορίθμου CBOW ή Skip-gram) θα πραγματοποιείται ανά τέσσερις λέξεις, δηλαδή «Είναι υπέροχο το καλοκαίρι», «υπέροχο το καλοκαίρι στην» κλπ.
6. Τέλος εκπαιδεύεται νευρωνικό δίκτυο πρόσω-τροφοδότησης με συνάρτηση ενεργοποίησης ιεραρχικής softmax ή αρνητικής δειγματοληψίας (Negative Sampling). Η ιεραρχική softmax αποδίδει καλύτερα όταν τα δεδομένα είναι χωρίς πολλές επαναλαμβανόμενες λέξεις, όμως είναι πιο αργό. Από την άλλη το Negative Sampling δουλεύει πιο γρήγορα και πιο καλά με επαναλαμβανόμενες λέξεις με την προϋπόθεση τα διανύσματα να μην είναι μεγάλων διαστάσεων (από 50 έως 100).

3.1.2 Υλοποίηση των CBOW και Skip-gram με νευρωνικά δίκτυα

Η υλοποίηση των μεθόδων Skip-gram και CBOW που αναφέρθηκαν παραπάνω με αρχιτεκτονικές νευρωνικών δικτύων φαίνεται στο παρακάτω σχήμα [19]:



Σχήμα 3.12 Οι αρχιτεκτονικές της κατανεμημένης αναπαράστασης

Με το CBOW προβλέπεται η λέξη, έχοντας δεδομένο τα συμφραζόμενα, ενώ με το Skip-gram συμβαίνει το αντίστροφο, προβλέπεται τα συμφραζόμενα, δεδομένης μιας λέξης. Η αρχιτεκτονική του Continuous Bag of Words (CBOW) είναι απλό μοντέλο ενός bag of words, δεδομένου τεσσάρων πλησιέστερων γειτόνων (δύο προηγούμενων και δύο επόμενων) και δεν λαμβάνει υπόψη την σειρά εμφάνισής τους.

Αναλυτικότερα με το k-skip-n-gram, νοείται ακολουθία μήκους n, στην οποία τα στοιχεία δεν απέχουν πάνω απόσταση k μεταξύ τους. Για παράδειγμα, η πρόταση «Είναι υπέροχο το καλοκαίρι στην Ελλάδα». Το 1-skip-2-gram θα αναλυθεί με συνδυασμούς των δύο λέξεων «Είναι υπέροχο», «υπέροχο το», «το καλοκαίρι», «καλοκαίρι στην», «στην Ελλάδα» συν τις ακολουθίες «Είναι το», «υπέροχο καλοκαίρι», «το στην», «καλοκαίρι Ελλάδα», δηλαδή προσπερνιέται μια λέξη (1-skip) και από εκείνα που έμειναν από δεξιά και αριστερά (τα προσπερασμένα) δημιουργούνται 2-gram.

3.1.3 Μερικές Παρατηρήσεις

Είναι αρκετά χρήσιμο να γνωρίζει κανείς ποιο αλγόριθμο ή αρχιτεκτονική θα χρειαστεί για την επίλυση του προβλήματος. Επίσης οι λεπτομέρειες, όπως ποια είναι η βέλτιστη παράμετρος και τι επηρεάζει, παίζουν σημαντικό ρόλο στο επιθυμητό αποτέλεσμα. Έτσι αξίζει να αναφερθούν οι εξής παρατηρήσεις :

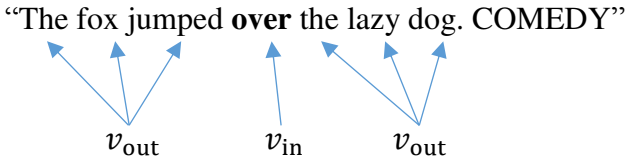
- Το CBOW τρέχει ταχύτερα, ενώ το Skip-gram αποδίδει καλύτερα, ειδικά με δεδομένα που περιέχουν αρκετά σπάνιες λέξεις
- Η ιεραρχική softmax επεξεργάζεται καλύτερα δεδομένα με σπάνιες λέξεις, ενώ η αρνητική δειγματοληψία (negative sampling) μοντελοποιεί αποτελεσματικότερα τις συχνές λέξεις
- Η παράμετρος subsampling συμβάλλει σημαντικά στην αποτελεσματικότητα του μοντέλου όταν οι τιμές της κυμαίνονται ανάμεσα στο 1e-3 και 1e-5
- Όσο περισσότερες είναι οι διαστάσεις του διανύσματος εμφύτευσης, τόσο καλύτερα
- Το μέγεθος του παράθυρου στο CBOW προτείνεται να ισούται με 5, ενώ στο Skip-gram το βέλτιστο είναι 10

Γενικά το μοντέλο CBOW είναι κατάλληλο για μεγάλο όγκο δεδομένων (πάνω από 100 000 000 λέξεις), διότι επεξεργάζεται καλύτερα τις συχνές λέξεις. Από την άλλη το Skip-gram αποδίδει αποτελεσματικότερα σε μικρότερα σύνολα κειμένων, αφού είναι πιο κατάλληλο για επεξεργασία σπάνιων λέξεων, αν και είναι πιο αργό.

3.2 Κατανεμημένη αναπαράσταση της παραγράφου (Doc2vec)

Η κατανεμημένη αναπαράσταση των παραγράφων είναι προέκταση της μεθόδου Word2vec. Μέσω της διαδικασίας Word2vec τα διανύσματα των λέξεων συμβάλλουν στην πρόβλεψη της επόμενης λέξης και παρόλο που η αρχικοποίηση των διανυσμάτων είναι τυχαία, εν τέλει οι λέξεις αποκτούν τέτοιες εμφύτευσεις ώστε να αντικατοπτρίζουν την σημασιολογία τους. Η μέθοδος Doc2vec λειτουργεί με παρόμοιο τρόπο, δηλαδή προβλέπει την λέξη, δεδομένου των συμφραζομένων που έχουν παρθεί από ολόκληρη την παράγραφο [15].

Με ανάλογο τρόπο το Doc2vec λειτουργεί πάνω στο προαναφερόμενο παράδειγμα, απλά προστίθεται άλλο ένα διάνυσμα, η λεγόμενη ετικέτα, που χαρακτηρίζει την παράγραφο. Για παράδειγμα, αν η κατηγοριοποίηση γίνεται με βάση το είδος της ταινίας, θα μπορούσε να ανήκει στην κωμωδία και η αναπαράσταση θα ήταν ως εξής:



Σχήμα 3.13 Απεικόνιση του παραθύρου doc2vec
Πηγή: https://events.yandex.ru/lib/talks/4137/

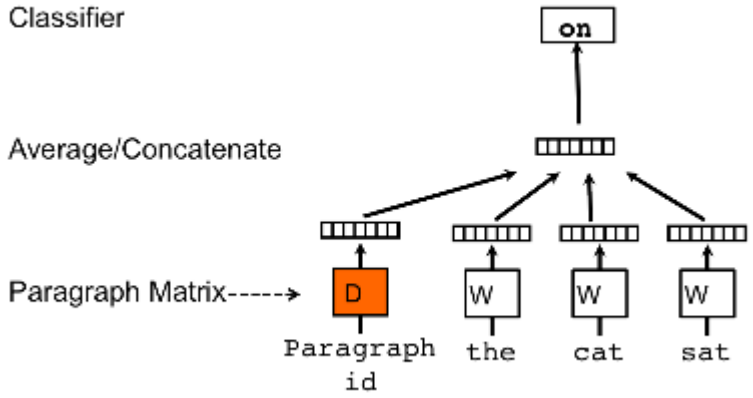
Επομένως, υπολογίζεται η πιθανότητα εμφάνισης της λέξης **jumped** δίπλα στην λέξη **over**, δεδομένου του είδους **comedy** στο οποίο ανήκει η πρόταση.

$$\mathrm{P}(v_{\text{out}} \mid v_{\text{in}}, \text{COMEDY}) = \mathrm{P}(v_{\text{jumped}} \mid v_{\text{over}}, v_{\text{comedy}})$$

3.2.1 Αρχιτεκτονική του Doc2vec

Η αρχιτεκτονική της κατανεμημένης αναπαράστασης της παραγράφου έχει ως εξής. Σε κάθε παράγραφο γίνεται ανάθεση ενός μοναδικού διανύσματος που αντιπροσωπεύει την παράγραφο αυτή στον πίνακα D. Το ίδιο συμβαίνει με τις λέξεις της παραγράφου, στις οποίες αναθέτονται μοναδικά διανύσματα του πίνακα W.

Έπειτα προβλέπεται η επόμενη λέξη μέσω του συνδυασμού του διανύσματος της παραγράφου και των λέξεων.



Σχήμα 3.14 Distributed Memory Model of Paragraph Vector – PV-DM

Στην ουσία, το διάνυσμα της παραγράφου μπορεί να θεωρηθεί το διάνυσμα μιας επιπλέον λέξης που απομνημονεύει αυτό που λείπει από τα συμφραζόμενα. Επίσης λειτουργεί και ως το θέμα της παραγράφου. Για αυτόν τον λόγο η διαδικασία αυτή ονομάζεται Κατανεμημένο Μνημονικό Μοντέλο του Διανύσματος της Παραγράφου (Distributed Memory Model of Paragraph Vector – PV-DM).

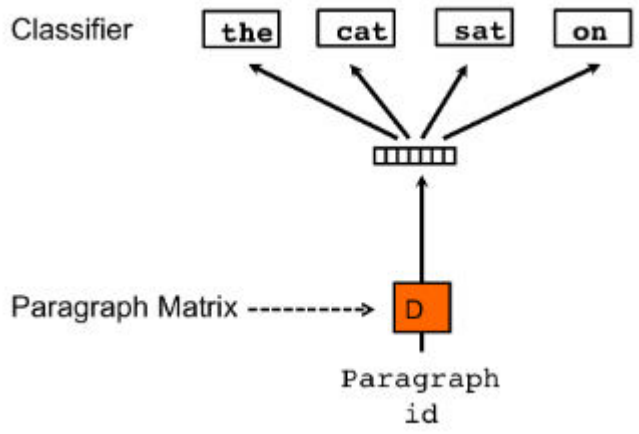
Σύμφωνα με αυτήν την διαδικασία, το παράθυρο μεταφέρεται στο μήκος της παραγράφου εξετάζοντας τα συμφραζόμενα. Ύστερα από την ανάλυση όλων των συνδυασμών των συμφραζομένων της ίδιας παραγράφου, παράγεται το διάνυσμά της. Το διάνυσμα της παραγράφου είναι μοναδικό και έχει παραχθεί από το περιεχόμενο της, ενώ τα διανύσματα των λέξεων έχουν διαμορφωθεί από όλες τις υπάρχουσες παραγράφους. Τα διανύσματα των λέξεων και των παραγράφων έχουν υπολογιστεί με βάση τον αλγόριθμο απότομης καθόδου. Σε κάθε βήμα του αλγορίθμου υπολογίζεται το σφάλμα από το δείγμα των συμφραζομένων που έχει επιλεχθεί τυχαία και έπειτα γίνεται ενημέρωση των παραμέτρων του μοντέλου βάση αυτού του σφάλματος.

Το πλήθος των παραμέτρων του μοντέλου είναι το αποτέλεσμα της σχέσης $N \times p + M \times q$, όπου N είναι το σύνολο των παραγράφων και M το σύνολο των λέξεων που υπάρχουν στο λεξικό. Ο στόχος είναι να προσδιοριστούν τα διανύσματα των παραγράφων, τέτοια ώστε ή κάθε παράγραφος να είναι p διαστάσεων και η κάθε

λέξη να είναι *q* διαστάσεων. Είναι προφανές πως ο αριθμός των παραμέτρων είναι μεγάλος, όταν οι παράγραφοι είναι πολλές.

Αξίζει να αναφερθεί μια σειρά από προτερήματα της κατανεμημένης αναπαράστασης των παραγράφων. Τα διανύσματα των παραγράφων δίνουν την δυνατότητα πιο αποτελεσματικής ανάλυσης δεδομένων που δεν διαθέτουν ετικέτες. Επιπλέον καλύπτουν τα μειονεκτήματα των παλιότερων μεθόδων, όπως η απώλεια της σημασιολογίας του κειμένου με την χρήση του bag-of-words. Με το Doc2vec, οι λέξεις που είναι κοντά εννοιολογικά έχουν μικρές αποστάσεις στον διανυσματικό χώρο, σε αντίθεση με τις λέξεις που δεν συσχετίζονται μεταξύ τους. Επίσης, λαμβάνεται υπόψη και το συντακτικό των μικρών τμημάτων της παραγράφου που δημιουργούνται με την βοήθεια του παραθύρου.

Μια άλλη αρχιτεκτονική της κατανεμημένης αναπαράστασης της παραγράφου είναι η Distributed Bag of Words (PV-DBOW), που στην πραγματικότητα είναι το αντίστροφο του προηγούμενου μοντέλου. Αντί λοιπόν να γίνεται η πρόβλεψη με βάση τον συνδυασμό των διανυσμάτων της παραγράφου και των λέξεων, προβλέπονται τυχαία οι λέξεις από την παράγραφο με την είσοδο του διανύσματός της, όπως φαίνεται στο παρακάτω σχήμα :



Σχήμα 3.15 Distributed Bag of Words (PV-DBOW)

Η αρχιτεκτονική αυτή είναι ανάλογη της Skip-gram από την κατανεμημένη αναπαράσταση των λέξεων. Συνήθως το διάνυσμα της παραγράφου προκύπτει από τον συνδυασμό δυο διανυσμάτων, ένα από το μοντέλο PV-DM και το άλλο από PV-DBOW, αν και το PV-DM δουλεύει ικανοποιητικά καλά και από μόνο του, όμως για καλύτερα αποτελέσματα προτιμάται ο συνδυασμός των δυο.

3.3 Gensim (GENerate SIMilarities)

Είναι ενδιαφέρον το γεγονός ότι δεν χρειάζεται να μπει κανείς στην διαδικασία να υλοποιήσει όλες τις παραπάνω διαδικασίες μόνος του, καθώς η λειτουργικότητα αυτή υπάρχει στο πακέτο Gensim της Python. Ακολουθώντας τις οδηγίες, διαπιστώνεται στην πράξη η ιδέα της κατανεμημένης αναπαράστασης των λέξεων και η δυνατότητα πράξεων μεταξύ τους και η διαδικασία ξεκινάει με την λήψη της βιβλιοθήκης [16].

Στον κώδικα δημιουργείται ένα μοντέλο το οποίο προσδιορίζει τον τρόπο με τον οποίο θα επεξεργαστούν τα δεδομένα :

```
num_features = 300    # Word vector dimensionality
min_word_count = 10    # Minimum word count
num_workers = 20         # Number of threads to run in parallel
context = 5                  # Context window size
downsampling = 1e-4  # Downsample setting for frequent words
neg_sample=5             #how many "noise words" should be drawn

model = Doc2Vec(min_count = min_word_count, window = context, size=num_features, sample = downsampling,
                negative=neg_sample, workers=num_workers)
```

Σχήμα 3.16 Απόσπασμα από τον κώδικα της κατανεμημένης αναπαράστασης

Η επεξήγηση των παραπάνω έχει ως εξής:

num_features - η διάσταση του παραγόμενου διανύσματος εμφύτευσης

min_word_count - η ελάχιστη συχνότητα εμφάνισης μιας λέξης

num_workers - αριθμός threads που συμβάλουν στην ταχύτερη επεξεργασία

context - η μέγιστη απόσταση μεταξύ της προβλεπόμενης και της επεξεργαζόμενης λέξης (μήκος παραθύρου)

downsampling - μείωση των πολύ συχνών λέξεων κατά την επεξεργασία

neg_sample - αριθμός των “αρνητικών” (τυχαίων) λέξεων που χρησιμοποιούνται ως αντιπαράδειγμα κατά την εκπαίδευση

Το τελικό μοντέλο αποθηκεύεται με την εντολή:

>>> model.save(model_file)

3.4 Έλεγχος αποτελεσμάτων

Έχοντας λοιπόν ένα αρχείο εισόδου .csv, τα δεδομένα του οποίου συλλέχθηκαν με την χρήση του API-VK και επεξεργάστηκαν με τον κώδικα του Doc2vec, ως αποτέλεσμα λαμβάνεται ένα αρχείο με τα σχετικά διανύσματα που αντιπροσωπεύουν το δοθέν λεξικό. Παρόλο που το μέγεθος του λεξικού που δημιουργήθηκε δεν είναι τόσο μεγάλο όσο θα μπορούσε να είναι αν η πηγή των δεδομένων θα ήταν η Wikipedia, ωστόσο τα αποτελέσματα είναι ικανοποιητικά. Αυτό διαπιστώνεται με την μέθοδο **most_similar** όπως φαίνεται παρακάτω, δηλαδή οι λέξεις με όμοια σημασιολογία βρίσκονται αντίστοιχα κοντά στον διανυσματικό χώρο.

Στο πρώτο παράδειγμα εξετάζεται η λέξη «Σαντορίνη», μιας και είναι ένας από τους πιο πολυσύχναστους τουριστικούς προορισμούς. Ανάμεσα στην πρώτη οχτάδα τυπώνονται οι λέξεις όπως το «νησί», η «Κρήτη», «Ρόδος» και «Κέρκυρα».

```
In [11]: sim = model.most_similar('санторини'.decode('utf8'), topn=8)

In [12]: for word, score in sim:
    print word.encode('utf8'), score
   ....:
остров 0.806892991066
да 0.79886329174
крит 0.790031909943
там 0.786679923534
это 0.785490751266
фото 0.773603498936
родос 0.767438411713
корфу 0.763833165169
```

Σχήμα 3.17 Απεικόνιση των αποτελεσμάτων του μοντέλου

Στο δεύτερο παράδειγμα ως είσοδος επιλέχθηκε η λέξη «κοινοβούλιο» και ανάμεσα στην πρώτη δεκάδα λέξεων εμφανίζονται: «ευρωπαϊκό», «κοινοβούλιο», «Ελλάδα», «Σύριζα» και «φόροι».

```
In [13]: sim = model.most_similar('парламент'.decode('utf8'), topn=10)

In [14]: for word, score in sim:
    print word.encode('utf8'), score
   ....:
#стиль 0.571134269238
европарламент 0.549922764301
#модно 0.542235732079
греции 0.533629894257
минфин 0.533601462841
подал 0.523532986641
сириза 0.519731104374
уже 0.519024431705
налоговых 0.517854511738
греция 0.516195654869
```

Σχήμα 3.18 Τρόπος εμφάνισης των πιο κοντινών λέξεων

Στο τρίτο παράδειγμα εφαρμόζονται οι πράξεις μεταξύ τον λέξεων, προστίθενται τα διανύσματα των λέξεων «Τσίπρας» και «Πούτιν», και αφαιρείται το διάνυσμα της «Ελλάδας». Στην ουσία εμφανίζονται εκείνες οι λέξεις που έχουν την μεγαλύτερη πιθανότητα να εμφανιστούν ανάμεσα στους δύο θετικούς όρους (Τσίπρας και Πούτιν), ενώ να μην αναφέρεται η συσχέτιση με τον αρνητικό όρο (Ελλάδα). Ανάμεσα στα αποτελέσματα εμφανίζονται οι λέξεις όπως «Ρωσία», «μετανάστες», και «πρόσφυγες».

```
In [12]: sim = model.most_similar(positive=['ципрас'.decode('utf8'),'путин'
.decode('utf8')], negative=['греция'.decode('utf8')],  topn=10)

In [13]: for word, score in sim:
             print word.encode('utf8'), score
   ....:
путина 0.63004052639
мигрантов 0.60681283474
или 0.597902178764
ес 0.593821704388
президента 0.589819788933
ципраса 0.589182496071
беженцев 0.58389377594
не 0.582327365875
на 0.58177947998
россии 0.581635594368
```

Σχήμα 3.19 Πράξεις μεταξύ των λέξεων

Στο τέταρτο παράδειγμα με τον ίδιο τρόπο, ως είσοδος, λαμβάνονται οι θετικοί όροι «νόστιμο» και «παραδοσιακό», ενώ αφαιρείται το διάνυσμα της λέξης «Ελλάδα». Τα

αποτελέσματα που εμφανίζονται είναι εξίσου ικανοποιητικά: «κουζίνα», «πρωτοχρονιάτικο», «κοκτέιλ», «γευσιγνωσία», «νόστιμο», «αρώματα».

```
In [21]: sim = model.most_similar(positive=['вкусное'.decode('utf8'),'традиционную'
.decode('utf8')], negative=['греция'.decode('utf8')],  topn=10)

In [22]: for word, score in sim:    print word.encode('utf8'), score
   ....:
кухню 0.24742987752
прятали 0.233922347426
новогодние 0.233106344938
знаменитые 0.226840361953
традицию 0.219061717391
коктейли 0.213752448559
дегустации 0.212831765413
пологий 0.210736900568
вкусную 0.208553045988
ароматы 0.204769432545
```

Σχήμα 3.20 Πράξεις μεταξύ των λέξεων

Ο πίνακας συντεταγμένων του διανύσματος μιας λέξης έχει την εξής μορφή:

```
In [60]: model['вкусное'.decode('utf8')]
Out[60]:
array([ 0.58408916, -0.10480614, -0.10813139,  0.19977023,  0.01027796,
        0.19071995, -0.16350724, -0.35989523, -0.05250763,  0.02044177,
        0.05574746,  0.18011697,  0.33794114,  0.03456053,  0.28042656,
        0.14012505,  0.1450167 , -0.34870377, -0.25784805, -0.05032127,
       -0.17608036,  0.08547288,  0.10661166,  0.11100371,  0.30298433,
        0.126243  ,  0.15206404,  0.14433263, -0.06922925, -0.08517742,
       -0.24068826,  0.16451287,  0.01796542,  0.05088358,  0.27489009,
        0.04454815, -0.1825401 , -0.1403251 ,  0.1219181 ,  0.04475064,
        0.05944694,  0.26653847,  0.10910118,  0.04401293,  0.11591685,
       -0.01309897,  0.14202155, -0.44351339,  0.07869779, -0.24097265,
       -0.02416862,  0.28664812,  0.18736161, -0.00737533, -0.13217671,
       -0.33114585, -0.34770903,  0.19903235,  0.18344602, -0.23847355,
        0.24059825, -0.10247122, -0.04958427, -0.4279483 , -0.32488075,
```

Σχήμα 3.21 Ο πίνακας συντεταγμένων του διανύσματος

Κάθε διάνυσμα αποτελείται από 300 διαστάσεις όπως έχει οριστεί από το μοντέλο:

```
        0.17562172, -0.33925733,  0.0334085 , -0.13694106,  0.12550627,
        0.08597753,  0.16849831,  0.11333947,  0.12444071,  0.06009077,
       -0.07866271,  0.2929599 , -0.24908198,  0.30755129,  0.08907419,
        0.05837474,  0.01310536, -0.00561402,  0.07047745, -0.0672408 ,
        0.2276244 ,  0.10697638,  0.05275571, -0.19265503, -0.06300903,
        0.42756554,  0.15088232,  0.18594988,  0.06222808,  0.00948735,
        0.02939254, -0.03582664, -0.14450949, -0.02858052,  0.04929309,
        0.15541154,  0.08364297, -0.08389155,  0.07719694,  0.24307835,
        0.1732714 ,  0.4857339 , -0.08423238,  0.25803959,  0.07452391,
        0.08926006, -0.27323326,  0.04566738,  0.18463667, -0.21813011], dtype=float32)

In [61]: len(model['вкусное'.decode('utf8')])
Out[61]: 300

In [62]:
```

Σχήμα 3.22 Ο πίνακας συντεταγμένων του διανύσματος

Τέλος, όπως έχει αναφερθεί η εφαρμογή του Doc2vec προσφέρει την δυνατότητα εμφάνισης ομοιότητας μεταξύ δύο λέξεων:

```
In [63]: model.similarity('вкусное'.decode('utf8'), 'традиционную'.decode('utf8'))
Out[63]: 0.23397064230307785

In [64]:
```

Σχήμα 3.23 Απεικόνιση ομοιότητας μεταξύ δύο λέξεων

4. Ομαδοποίηση των δεδομένων

Ομαδοποίηση είναι η διαδικασία κατά την οποία τα δεδομένα κατατάσσονται σε σημασιολογικά σύμφωνες ομάδες (clusters) με βάση κάποιο μέτρο ομοιότητας. Δηλαδή δεδομένα που ανήκουν στην ίδια ομάδα να είναι όμοια μεταξύ τους, ενώ δεδομένα από διαφορετικές ομάδες να είναι ανόμοια. Είναι ένα πρόβλημα μη επιβλεπόμενης μηχανικής μάθησης (unsupervised machine learning) , που σημαίνει πως η δομή των δεδομένων πρέπει να ανιχνευτεί χωρίς να είναι διαθέσιμη κάποια ετικέτα για το σε ποια κατηγορία ανήκουν. Πρόκειται για μια υποκειμενική διαδικασία, καθώς το ίδιο σύνολο δεδομένων πρέπει να διαχωριστεί διαφορετικά, ανάλογα την εφαρμογή. Σε αυτή την υποκειμενικότητα έγκειται και η δυσκολία της ομαδοποίησης, καθώς ένας αλγόριθμος ή μία συγκεκριμένη προσέγγιση δεν επαρκούν για να λύσουν κάθε πρόβλημα ομαδοποίησης. Ένας από τους πιο κλασικούς αλγόριθμους ομαδοποίησης αποτελεί ο αλγόριθμος K-Means, ο οποίος αναθέτει τα στοιχεία του διανυσματικού χώρου σε προεπιλεγμένο αριθμό των ομάδων $\boldsymbol{k}$. Είναι μια επαναληπτική διαδικασία με τα εξής βήματα [3]:

1. Επιλέγεται ο αριθμός των ομάδων $\boldsymbol{k}$
2. Η αρχικοποίηση των $\boldsymbol{k}$ κέντρων γίνεται με τυχαίο τρόπο
3. Για κάθε σημείο επιλέγεται το κοντινότερο κέντρο με βάση την ευκλείδεια απόσταση, έτσι τα διανύσματα που είναι πιο κοντά στο κέντρο αυτό, ανατίθενται στο συγκεκριμένο κέντρο
4. Υπολογίζονται εκ νέου τα κέντρα, δηλαδή τα κέντρα βάρους των ομάδων, καθώς τα νέα κέντρα είναι τα διανύσματα που έχουν προκύψει από τον μέσο όρο των σημείων που έχουν ανατεθεί σε αυτά

Με την βοήθεια της ευκλείδειας απόστασης υπολογίζονται οι αποστάσεις των σημείων από τα κέντρα:

$$\mathrm{d(x,y)} = \|\mathrm{x} - \mathrm{y}\| = \sqrt{\sum_{p=1}^{n}(x_p - y_p)^2}\ , \quad \text{όπου x, y} \in R^n$$

Με σημεία $(x^{(1)}, x^{(2)}, \dots, x^{(m)}) \in R^n$ και S = (S_1, S_2, … , S_k), ο αλγόριθμος διασπά τις $\boldsymbol{m}$ παρατηρήσεις σε $\boldsymbol{k}$ ομάδες (k≤m), ώστε να ελαχιστοποιείται η συνολική τετραγωνική απόκλιση των σημείων από τα κέντρα [2]:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - \mu_j||^2$$

όπου $||x_i^{(j)} - \mu_j||^2$, είναι ένα μέτρο απόστασης που χρησιμοποιείται για να μετρά την απόσταση κάθε στοιχείου x_i από το centroid μ_j της κάθε ομάδας, και n ο αριθμός των στοιχείων του συνόλου δεδομένων. Ένας αντιπροσωπευτικός διαμεριστικός αλγόριθμος, ο οποίος χρησιμοποιεί αυτή τη συνάρτηση και θα αναλυθεί στην συνέχεια είναι ο K-Means.

Τα διανύσματα που έχουν χωριστεί σε $\boldsymbol{k}$ ομάδες, προσδιορίζουν το κέντρο της ίδιας ομάδας. Το κάθε διάνυσμα ανήκει στο κέντρο που βρίσκεται πιο κοντά σε αυτό, με την προϋπόθεση το κάθε σημείο να αντιστοιχίζεται σε μια μόνο ομάδα. Σε κάθε επανάληψη το κέντρο υπολογίζεται ως το μέσο των σημείων που έχουν ανατεθεί στην ομάδα [18]:

$\mu_j = \frac{1}{n_j} \sum_{x^{(j)} \in n_i} x^{(j)}$, όπου n_j ο αριθμός των σημείων που έχουν ανατεθεί στην ομάδα j

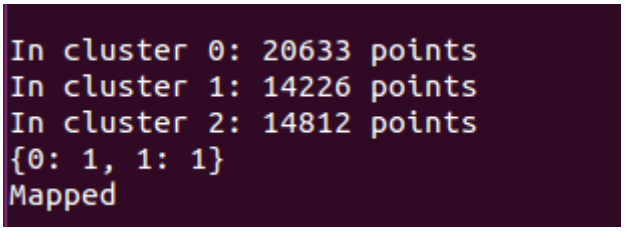
Στην συνέχεια επαναλαμβάνεται το τρίτο και το τέταρτο βήμα, ενώ κατά την επανάληψη τα κέντρα και τα όρια των ομάδων μετακινούνται. Εν τέλει, ελαχιστοποιούνται οι αποστάσεις μεταξύ των σημείων της ομάδας από τα κέντα στα οποία ανατίθενται. Ο αλγόριθμος τερματίζεται όταν τα κέντρα μένουν στην θέση που ήταν και στην προηγούμενη επανάληψη με το ίδιο σύνολο σημείων σε κάθε ομάδα:

$$\mu_i^{\beta\acute{\eta}\mu\alpha\ t} = \mu_i^{\beta\acute{\eta}\mu\alpha\ t+1}$$

Συνήθως η διαδικασία απαιτεί δεκάδες επαναλήψεις ώσπου να σταθεροποιηθούν τα κέντρα. Πλεονεκτήματα του αλγορίθμου είναι η απλότητα της διαδικασίας. Το μειονέκτημα από την άλλη είναι η αβεβαιότητα της επιλογής των αρχικών κέντρων, του αριθμού των ομάδων, κάτι που απαιτεί ουσιαστική γνώση των δεδομένων, αλλά και του υπολογιστικού κόστους ώστε να υπολογιστούν όλες οι απαιτούμενες αποστάσεις.

4.1 Πειραματικά αποτελέσματα

Οι 49.671 δημοσιεύσεις που μετατράπηκαν σε διανύσματα, ομαδοποιήθηκαν σε 3 ομάδες (*k=3*) με τον αλγόριθμο K-Means. Η επιλογή των 3 ομάδων στηρίχτηκε στα κυρίως θέματα των σελίδων του VK (τουρισμός, κοινωνικά και γαστρονομία). Όπως φαίνεται στο παρακάτω σχήμα, 20.633 κείμενα ανατέθηκαν με βάση τον αλγόριθμο K-Means και βέλτιστη αρχικοποίηση των κέντρων με τον αλγόριθμο K-Means++ στην πρώτη ομάδα, 14.226 στην δεύτερη και 14.812 στην τρίτη ομάδα.



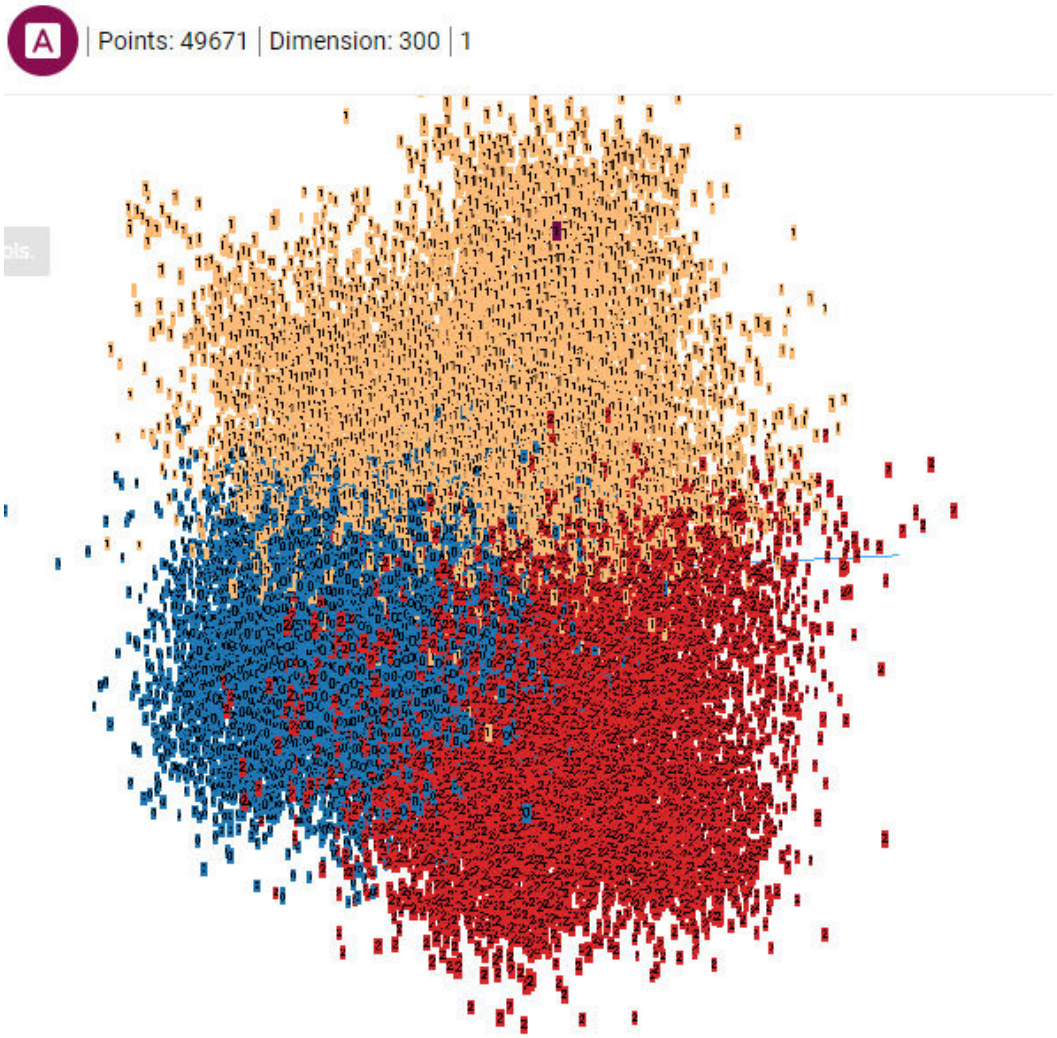
Σχήμα 4.1 Ομαδοποίηση των σημείων με k=3

Για λόγους οπτικοποίησης χρησιμοποιήθηκε το εργαλείο Embedding Projector της TensorFlow, το οποίο διαβάζει τις εμφυτεύσεις από το επιλεγόμενο αρχείο και τις προβάλει σε τρεις διαστάσεις με την βοήθεια της Ανάλυσης Κύριων Συνιστωσών (Principal Component Analysis). Η μέθοδος PCA είναι μια γραμμική τεχνική συμπίεσης δεδομένων που επαναπροσδιορίζει τις συντεταγμένες των αρχικών διανυσμάτων σε ένα νέο σύστημα συντεταγμένων μικρότερων διαστάσεων, το οποίο είναι πιο κατανοητό ως προς την ανάλυση και παρατήρηση. Με την βοήθεια του αντίστροφου μετασχηματισμού επιλέγεται ο άξονας με την μεγαλύτερη διασπορά και έτσι όταν προβληθούν τα δεδομένα πάνω στην γραμμή, τα σημεία αυτά θα έχουν την μέγιστη διασπορά. Οι κύριες συνιστώσες διατηρούν τις διακυμάνσεις των δεδομένων και η συνολική τους ποσότητα είναι ίδια με την αρχική. Η PCA μέθοδος μπορεί να αναλυθεί στα παρακάτω βήματα:

1. Υπολογίζεται ο μέσος όρος των στοιχείων κάθε στήλης και ύστερα αφαιρείται από κάθε στοιχείο της αντίστοιχης στήλης. Η μέση τιμή του παραγόμενου συνόλου δεδομένων είναι προφανώς ίση με το μηδέν.
2. Υπολογίζεται ο πίνακας διακύμανσης και συνδιακύμανσης. Το άθροισμα των στοιχείων της διαγώνιου εκφράζει την ολική διακύμανση.

3. Υπολογίζονται οι ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα συνδιακύμανσης
4. Επιλέγεται το ιδιοδιάνυσμα με την υψηλότερη ιδιοτιμή, το οποίο θα αποτελέσει την πρώτη κύρια συνιστώσα και περιέχει περισσότερες πληροφορίες σε σχέση με την δεύτερη.
5. Συλλέγονται τα τελικά δεδομένα, τα οποία προκύπτουν από τον πολλαπλασιασμό του ανάστροφου του χαρακτηριστικού διανύσματος με το αρχικό σύνολο δεδομένων
6. Έπειτα, επαναφέρονται τα δεδομένα στο αρχικό τους στάδιο και υπολογίζεται το σφάλμα συμπίεσης

Η αναπαράσταση των κειμένων μετατράπηκαν σε διανύσματα 300 διαστάσεων και έχουν μετασχηματιστεί σε διανύσματα τριών διαστάσεων φαίνεται στο παρακάτω σχήμα:



Σχήμα 4.2 Απεικόνιση της ομαδοποίησης των σημείων με k=3

5. Θεματική μοντελοποίηση

Ένας τρόπος εξαγωγής πληροφορίας από μια συλλογή κειμένων είναι η θεματική μοντελοποίση (topic modeling). Η θεματική μοντελοποίση δημιουργεί ένα μοντέλο το οποίο ορίζει σε ποιο θέμα (topic) αντιστοιχεί το κάθε έγγραφο. Το πιθανολογικό θεματικό μοντέλο (ΠΘΜ) περιγράφει το κάθε θέμα ως διακριτή κατανομή των όρων/λέξεων, ενώ το κάθε έγγραφο αναπαρίσταται ως διακριτή κατανομή σε πολλαπλά θέματα. Μιας και ένα έγγραφο ή ένας όρος μπορεί να αντιστοιχεί ταυτόχρονα σε πολλά θέματα με διαφορετικές πιθανότητες, το ΠΘΜ πραγματοποιεί την λεγόμενη «ήπια» ομαδοποίηση. Με αυτόν τον τρόπο επιλύονται τα προβλήματα συνωνυμίας των όρων που προκύπτουν από την «αυστηρή» ομαδοποίηση, όταν δηλαδή το έγγραφο ή ο όρος ανήκει σε μία συγκεκριμένη θεματολογία.

Έχει διαπιστωθεί πως τα μοντέλα με κρυμμένες (latent) μεταβλητές είναι τα πιο αποδοτικά στην κατανόηση του περιεχόμενου ενός κειμένου. Τα μοντέλα αυτά επιλύουν διάφορα προβλήματα, όπως, ομαδοποίηση εγγράφων, εύρεση όμοιων εγγράφων, πολύγλωσση αναζήτηση, εντοπισμός των λέξεων-κλειδιών, εύρεση της εξάρτησης μεταξύ των όρων κτλπ.

Για παράδειγμα, στην προσπάθεια να βρεθούν οι δημοσιεύσεις που να έχουν σχέση με «Ελλάδα», «capital control» και «Τσίπρας», το θεματικό μοντέλο είναι ικανό να εντοπίσει τα πιο συναφή κείμενα που έχουν σχέση με τις οικονομικές εξελίξεις, όσον αφορά στην τήρηση των capital controls στην Ελλάδα κατά την θητεία του πρωθυπουργού Α. Τσίπρα. Το αξιοσημείωτο είναι πως δεν είναι απαραίτητη η παρουσία των συγκεκριμένων λέξεων-κλειδιών στα κείμενα αυτά. Τα πιο συναφή άρθρα που πιθανόν να προκύψουν μπορεί να μην περιέχουν την λέξη «Τσίπρας», παρόλα αυτά ο συνδυασμός των λέξεων που έχουν δοθεί, δημιουργούν μια ή περισσότερες κρυμμένες μεταβλητές που έχουν σχέση με το γενικό ζητούμενο θέμα.

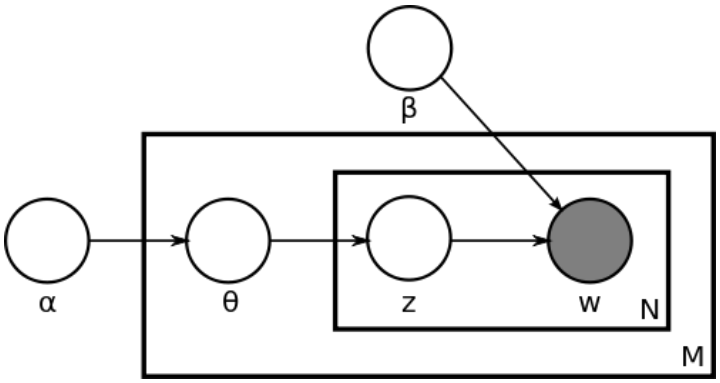
Από την άλλη πλευρά υπάρχουν τεχνικές, βασιζόμενες στον υπολογισμό συχνότητας εμφάνισης των όρων (term-frequency inverse-document-frequency TF-IDF). Στην προκειμένη περίπτωση ισχύει το αντίστροφο από την προαναφερόμενη μέθοδο, δηλαδή ο αλγόριθμος επικεντρώνεται στους συνδυασμούς μεταξύ των λέξεων-κλειδιών που πρέπει να εμπεριέχονται στα κείμενα.

Ένα ιδανικό μοντέλο καλείται να συνδυάσει και τους δυο παραπάνω τρόπους, έτσι ώστε τα τελικά αποτελέσματα να είναι τα βέλτιστα δυνατά. Μια λύση μπορεί να αποτελέσει η LDA, όπως περιγράφεται στην επόμενη ενότητα.

5.1 Latent Dirichlet Allocation (LDA)

Μια από τις πιο κατάλληλες και αποτελεσματικές μεθοδολογίες για θεματική ομαδοποίηση είναι η Latent Dirichlet Allocation (LDA) [24]. Πρόκειται για έναν αλγόριθμο που επιλύει προβλήματα ανάλυσης δεδομένων δημιουργώντας πιθανολογικά μοντέλα μεγάλου όγκου κειμένων [27]. Αυτό που την κάνει να διαφέρει από άλλους τρόπους θεματικής ομαδοποίησης είναι ότι λαμβάνει υπόψιν το γεγονός ότι ένα έγγραφο μπορεί να περιέχει πάνω από ένα θέμα. Το LDA είναι ένα ιεραρχικό μοντέλο, βασιζόμενο στην θεωρία του Bayes και αποτελείται από δύο επίπεδα [28]:

- Το πρώτο είναι ένα μίγμα, συστατικά του οποίου αντιστοιχούν στα θέματα
- Το δεύτερο αντιπροσωπεύεται από μια μεταβλητή της κατανομής Dirichlet που χωρίζει το έγγραφο σε θέματα



Σχήμα 5.1 Το θεματικό μοντέλο δημιουργήθηκε το 2003 από τον David Blei, Andrew Ng και τον Michael Jordan

Για την πλήρη κατανόηση ενός πολύπλοκου μοντέλου, αρκεί να γίνει αντιληπτή η διαδικασία επεξεργασίας ενός καινούργιου έγγραφου κατά βήμα [17]:

- Ορίζεται το μήκος του έγγραφου N (αριθμός λέξεων)
- Ορίζεται το διάνυσμα $\theta \sim (\alpha)$, το οποίο εκφράζει την περιεκτικότητα του θέματος μέσα στο έγγραφο
- Για κάθε λέξη w :
    1. Επιλέγεται θέμα z_n με βάση την κατανομή *Mult(θ)*
    2. Επιλέγεται λέξη $w_n \sim p(w_n|z_n, \beta)$ με πιθανότητες δοσμένες από το β

Για την απλοποίηση της διαδικασίας ορίζεται ο αριθμός των θεμάτων k και θεωρείται πως το β είναι ένα σύνολο παραμέτρων $\beta_{ij} = p(w^j=1|z^i=1)$ που πρέπει να υπολογιστούν, χωρίς να λαμβάνεται υπόψη το N. Συνεπώς, ο τύπος του μοντέλου είναι [28]:

$$p(\theta...N \mid \alpha,\beta) = p(N...\xi)\, p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n|\theta)\, p(w_n|z_n,\beta)$$

Έτσι, η ομάδα δεν υπολογίζεται μια φορά και ύστερα προσδιορίζονται οι λέξεις, αλλά για την κάθε λέξη επιλέγεται το θέμα από την κατανομή θ και έπειτα προσδιορίζεται η λέξη σε σχέση με αυτό το θέμα.

Με την ολοκλήρωση της εκπαίδευσης του μοντέλου, επιστρέφονται διανύσματα $\boldsymbol{\theta}$ που δηλώνουν τον τρόπο που είναι κατανεμημένα τα θέματα σε κάθε έγγραφο, ενώ η κατανομή $\boldsymbol{\beta}$ εμφανίζει τις λέξεις που έχουν την μεγαλύτερη πιθανότητα να εμφανιστούν σε κάθε θέμα. Με αυτόν τον τρόπο, από τα αποτελέσματα του LDA λαμβάνεται μια λίστα θεμάτων που εμπεριέχονται σε κάθε κείμενο και για κάθε θέμα μια σειρά από λέξεις που το χαρακτηρίζουν. Αξιοσημείωτο το γεγονός ότι η διαδικασία γίνεται χωρίς επιτήρηση (unsupervised learning), δηλαδή η συλλογή μπορεί να αποτελείται από κείμενα χωρίς ετικέτες (labels).

Με απλούστερα λόγια, έχοντας μια συλλογή κειμένων, το κάθε ένα από αυτά εκλαμβάνεται σαν ένα bag of words (BOW). Το ζητούμενο είναι να προσδιοριστούν τα θέματα που αναφέρονται στο κείμενο, ενώ είναι άγνωστα τα πιθανά θέματα και από ποιες και πόσες λέξεις περιγράφονται. Στην ουσία αυτό ακριβώς επιλύει το πιθανολογικό μοντέλο, βρίσκοντας πρώτα το σύνολο των θεμάτων που υπάρχει στην δεδομένη συλλογή κειμένων. Επίσης σε κάθε θέμα αντιστοιχεί μια κατανομή η οποία προσδιορίζει τις πιο συχνά εμφανιζόμενες λέξεις που αποτελούν τις λέξεις-κλειδιά για τα προτεινόμενα θέματα [26].

Στην προκειμένη περίπτωση ο LDA αλγόριθμος διαβάζει το αρχείο .csv με το υπάρχον λεξικό και εντοπίζει τα θέματα στα ήδη ομαδοποιημένα δεδομένα. Ενδεικτικά ορίστηκε ο αριθμός των ομάδων ίσο με δέκα, που σημαίνει ότι η κατανομή Dirichlet χωρίζει τις τρεις ομάδες σε δέκα υποομάδες. Είναι γνωστό εκ των προτέρων ότι τα άρθρα και οι λέξεις που δεν φέρνουν σημαντική πληροφορία θα αποτελέσουν ξεχωριστά θέματα και δεν θα λαμβάνονται υπόψη.

5.1.1 Εργαλεία απεικόνισης των LDA αποτελεσμάτων

Η LDA αποτελεί μια καλή επιλογή για την ανάλυση των δημοσιεύσεων του VK που συλλέχθηκαν. Είναι εφικτός ο εντοπισμός των θεμάτων στις ομάδες κειμένων που προέκυψαν από τον αλγόριθμο K-Means. Όμως το πρόβλημα είναι πως τα αποτελέσματα που προκύπτουν από το μοντέλο LDA δεν είναι πάντα εύκολα ερμηνεύσιμα από τον άνθρωπο. Αν και τα τελευταία χρόνια έχουν δημιουργηθεί αρκετά εργαλεία που έχουν σκοπό να προσφέρουν όσο το δυνατόν πιο κατανοητή απεικόνιση των θεματικών μοντέλων, τα περισσότερα από αυτά χρησιμοποιούν λίστες των πιο συχνών όρων που εμπεριέχονται σε ένα θέμα.

Το 2012 αναπτύχθηκε ένα πολύ χρήσιμο εργαλείο το «Termite», οι δημιουργοί του οποίου παρουσίασαν δύο καινούργια μέτρα που βελτιστοποιούν σημαντικά την απεικόνιση των θεμάτων [10]. Ο λόγος γίνεται για «saliency» και «distinctiveness» των όρων, δηλαδή πόση πληροφορία μεταφέρει ο όρος σε ένα θέμα. Η «υπεροχή» χρησιμοποιείται για την ταξινόμηση και το φιλτράρισμα των όρων. Με την εμφάνιση των πιο χαρακτηριστικών όρων είναι εφικτή η πιο γρήγορη εκτίμηση και σύγκριση των θεμάτων. Ο στόχος του «Termite» είναι να προσφέρει αντικειμενική εκτίμηση για τις κατανομές των όρων που είναι συνδεδεμένες με τα LDA θέματα και να αξιολογεί την ποιότητα του κάθε θέματος μεμονωμένα, αλλά και την ποιότητα όλων των θεμάτων συνολικά.

Η χρήση όλων των λέξεων στην κατανομή των όρων κάθε θέματος δεν είναι ιδιαίτερα περιγραφική. Για τον λόγο αυτό είναι σημαντικό το φίλτρο του «Termite» ώστε να εμφανίζονται εν τέλει οι λέξεις με την μεγαλύτερη πιθανότητα ή αλλιώς οι όροι που «υπερέχουν». Ο υπολογισμός αυτών γίνεται ως εξής:

Για κάθε λέξη $\boldsymbol{w}$ βρίσκεται η πιθανότητα $\boldsymbol{P(T \mid w)}$ που αντιπροσωπεύει την παρατηρούμενη λέξη $\boldsymbol{w}$ να έχει παραχθεί από κρυμμένο θέμα $\boldsymbol{T}$. Επίσης υπολογίζεται η οριακή πιθανότητα $\boldsymbol{P(T)}$, που αντιστοιχεί σε μια τυχαία λέξη $\boldsymbol{w'}$ που έχει παραχθεί από το θέμα $\boldsymbol{T}$. Η ιδιαιτερότητα (distinctiveness) της λέξης $\boldsymbol{w}$ υπολογίζεται με βάση τον τύπο:

$$\text{distinctiveness (w)} = \sum_T P(T|w) log \frac{P(T|w)}{P(T)}$$

Αυτός ο τύπος περιγράφει την συνεισφορά του όρου ***w*** στο παραγόμενο θέμα, σε σχέση με μια τυχαία επιλεγμένη λέξη ***w´***. Για παράδειγμα, αν μια λέξη εμφανίζεται σε όλα τα θέματα, προκύπτει ένα συμπέρασμα για το θεματικό μίγμα του εγγράφου, δηλαδή ο ορισμός θα αποκτήσει χαμηλό βαθμό «ιδιαιτερότητας». Η υπεροχή (saliency) της λέξης υπολογίζεται με τον εξής τρόπο:

$$Saliency(w) = P(w)^{x} distinctiveness(w)$$

Στην πραγματικότητα, αυτό το φιλτράρισμα επιδιώκει την γρήγορη ομαδοποίηση και αποσαφήνιση των θεμάτων, ξεχωρίζοντας τις λέξεις με μεγάλη πιθανότητα εμφάνισης (που πολλές φορές είναι γενικές λέξεις που δεν προσφέρουν κάποια διαφορά στο κείμενο) από τις λέξεις που ουσιαστικά χαρακτηρίζουν ένα θέμα. Με αυτόν τον τρόπο ξεχωρίζονται οι σημαντικές λέξεις από λέξεις χωρίς ιδιαίτερο πληροφοριακό περιεχόμενο.

5.1.2 Απεικόνιση LDAvis

Το «LDAvis» αποτελεί ένα αποτελεσματικό εργαλείο απεικόνισης των αποτελεσμάτων της LDA. Ο στόχος αυτού είναι να απαντήσει σε τρία βασικά ερωτήματα:

1. Ποια είναι η σημασία του κάθε θέματος
2. Πόσο κοινό είναι το κάθε θέμα
3. Ποιες είναι οι σχέσεις μεταξύ των θεμάτων

Η απεικόνιση των αποτελεσμάτων αποτελείται από δύο βασικά μέρη. Στην αριστερή μεριά παρουσιάζεται η γενική εικόνα του θεματικού μοντέλου, όπου απαντώνται οι ερωτήσεις 2 και 3. Οι κύκλοι στο συγκεκριμένο σχέδιο αντιπροσωπεύουν τα θέματα με τα κέντρα ορισμένα σύμφωνα με την απόσταση μεταξύ των θεμάτων και προσαρμοσμένα σε δύο διαστάσεις. Έπειτα, το μέγεθος του κύκλου εξαρτάται από την συνολική συχνότητα εμφάνισης του εγγράφου με φθίνουσα ταξινόμηση ως προς την συχνότητα αυτή.

Από την άλλη, το δεξί μέρος απεικονίζει ένα οριζόντιο ιστόγραμμα, οι στήλες του οποίου αντιστοιχούν στους πιο αντιπροσωπευτικούς όρους, που ερμηνεύουν το επιλεγόμενο θέμα στα αριστερά. Με αυτόν τον τρόπο απαντάται η πρώτη ερώτηση. Η διχρωμία των στηλών δείχνει την σχέση της συχνότητας εμφάνισης της

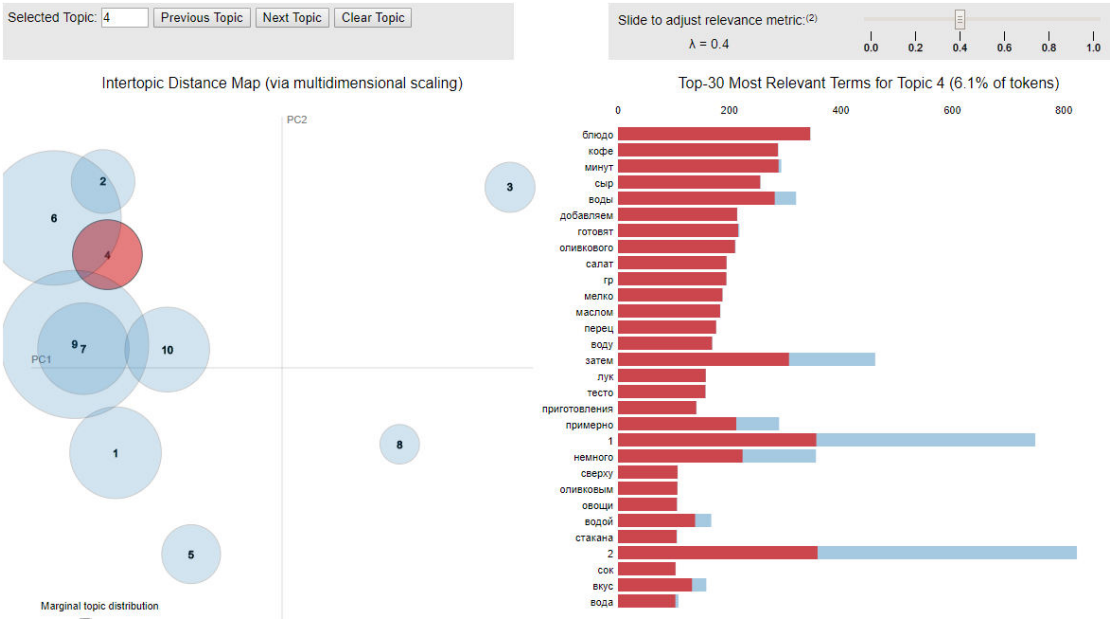
συγκεκριμένης λέξης σε όλη την έκταση της συλλογής των κειμένων, με την συχνότητα εμφάνισης μέσα στο επιλεγμένο θέμα. Με την αλληλεξάρτηση των δύο μελών της απεικόνισης φαίνεται η σύνδεση μεταξύ των θεμάτων και των πιο σημαντικών όρων. Αυτό επιτρέπει στον χρήστη να δουλεύει με μεγάλο όγκο δεδομένων, έχοντας πλήρη εικόνα του θεματικού μοντέλου.

Πέρα από τον υπολογισμό της υπεροχής (saliency) των λέξεων, το σύστημα δίνει βάση στην «σχετικότητα» μεταξύ της λέξης και του θέματος [4]. Με το $\boldsymbol{\varphi_{kw}}$ να ορίζει την πιθανότητα της λέξης $\boldsymbol{w} \in \{\mathbf{1}, \ldots, \boldsymbol{V}\}$ για το θέμα $\boldsymbol{k} \in \{\mathbf{1}, \ldots, \boldsymbol{K}\}$, όπου το $\boldsymbol{V}$ υποδηλώνει τον αριθμό των λέξεων στο λεξικό, και το $\boldsymbol{p_w}$ να ορίζει την οριακή πιθανότητα της λέξης $\boldsymbol{w}$ στην συλλογή των κειμένων, χτίζεται ο παρακάτω τύπος. Η «σχετικότητα» που υπάρχει μεταξύ της λέξης $\boldsymbol{w}$ και του θέματος $\boldsymbol{k}$ ορίζεται από την παράμετρο $\boldsymbol{\lambda}$, όπου $\mathbf{0} \leq \boldsymbol{\lambda} \leq \mathbf{1}$.

$$r(w,k \mid \lambda)\ \lambda \log(\varphi_{kw}) + (1- \lambda) \log(\frac{\varphi_{kw}}{p_w})$$

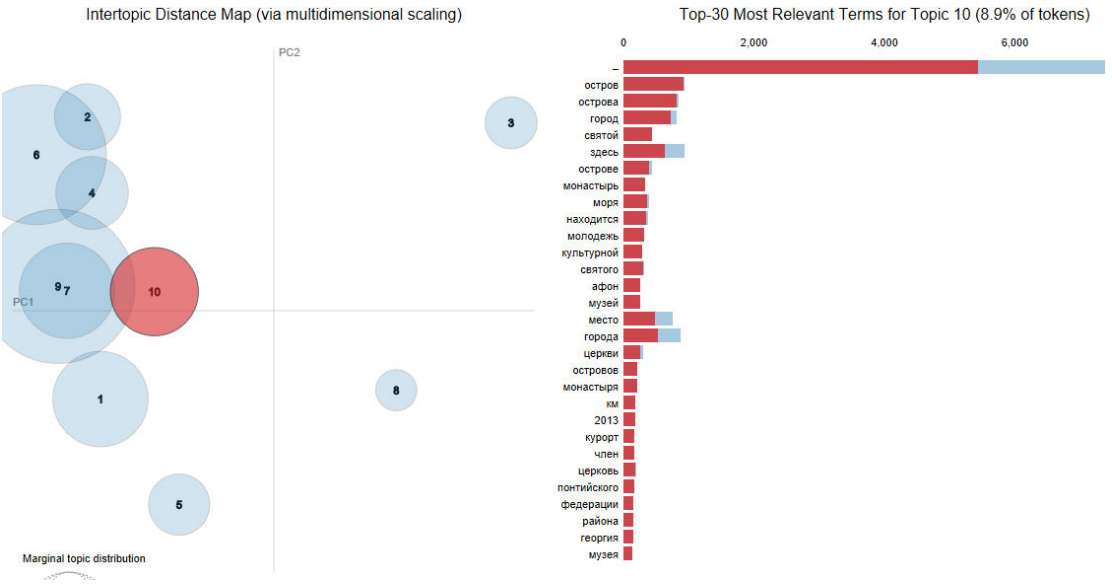
Το $\boldsymbol{\lambda}$ εκφράζει το βάρος που δίνεται στην πιθανότητα της λέξης $\boldsymbol{w}$, όταν αυτή εμπεριέχεται στο θέμα $\boldsymbol{k}$. Με το $\boldsymbol{\lambda} = \mathbf{1}$, οι ορισμοί ταξινομούνται με φθίνουσα σειρά ως προς την πιθανότητα εμφάνισης στο συγκεκριμένο θέμα, ενώ όταν το $\boldsymbol{\lambda} = \mathbf{0}$, η ταξινόμηση πραγματοποιείται με την αύξουσα σειρά.

Έτσι, με την βοήθεια του LDAvis αναλύονται οι τρεις ομάδες που έχουν προκύψει από τον αλγόριθμο K-Means σε 10 θέματα. Για παράδειγμα το τέταρτο θέμα της πρώτης ομάδας εκφράζεται αποκλειστικά από γαστρονομικούς όρους:



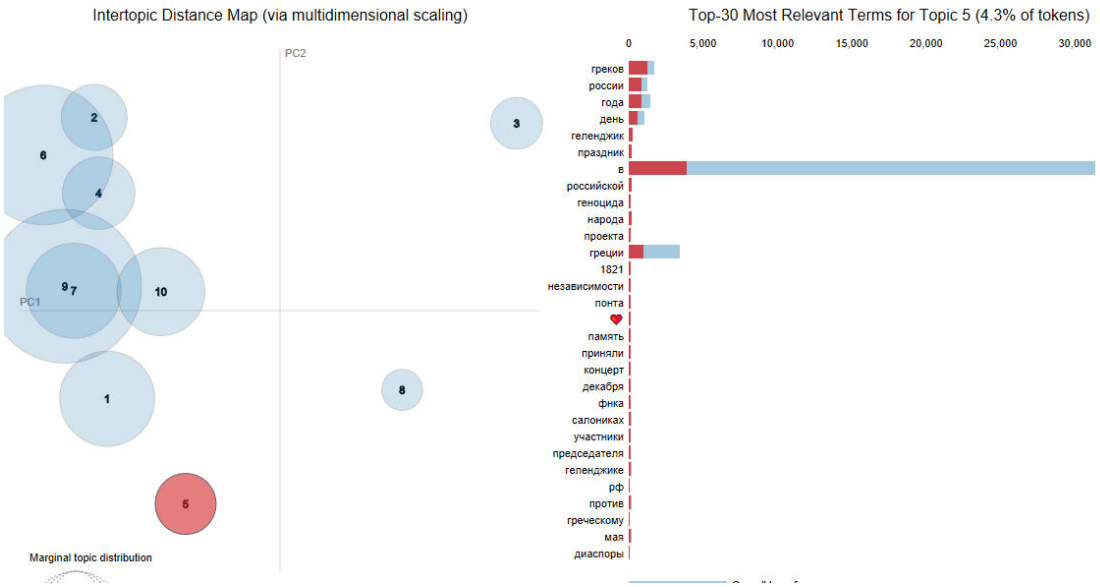
Σχήμα 5.2 Απεικόνιση του τέταρτου θέματος της πρώτης ομάδας

Το δέκατο θέμα αναφέρεται στα θρησκευτικά (μοναστήρι, μουσείο, Άγιο Όρος, εκκλησία, άγιος, Γεώργιος, ποντιακός).



Σχήμα 5.3 Απεικόνιση του δέκατου θέματος της πρώτης ομάδας

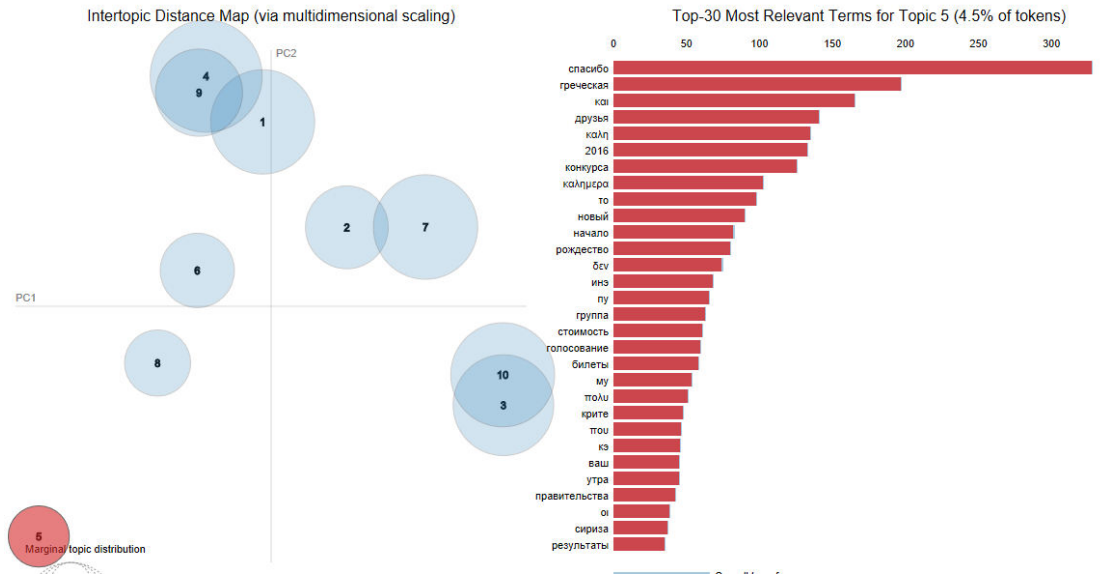
Οι λέξεις που σχετίζονται με την ιστορία βρίσκονται στο πέμπτο θέμα (Έλληνες, Ρωσία, γιορτή, λαός, Πόντος, 1821, γενοκτονία, μνήμη, ομογένεια).



Σχήμα 5.4 Απεικόνιση του πέμπτου θέματος της πρώτης ομάδας

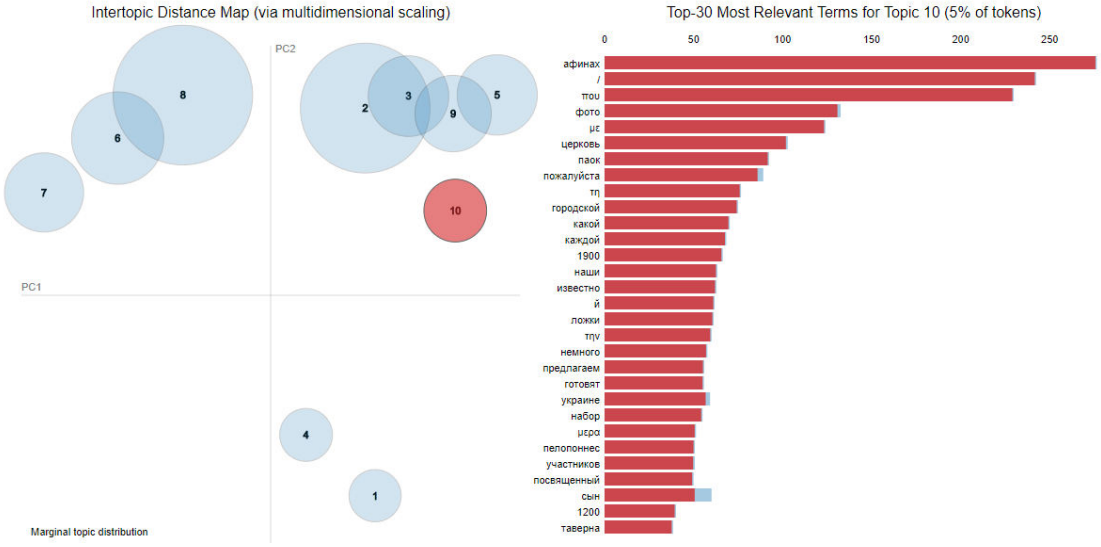
Η πρώτη ομάδα εμπεριέχει κατά βάση θέματα σχετιζόμενα με την γαστρονομία, θρησκευτικά και ιστορία.

Στην δεύτερη ομάδα παρατηρείται μεγαλύτερη έμφαση στους πολιτικούς όρους, συγκεκριμένα στο πέμπτο και στο όγδοο θέμα περιλαμβάνονται οι λέξεις όπως εκλογές, κυβέρνηση, Σύριζα, αποτελέσματα, βουλή, ρωσικό, ελληνικό, ΗΠΑ, Σαββίδης, ΜΜΕ.



Σχήμα 5.6 Απεικόνιση του όγδοου θέματος της δεύτερης ομάδας

Τέλος, στην τρίτη ομάδα παρατηρούνται οι όροι που φαίνεται να προέρχονται από δημοσιεύσεις σχετικές με τον τουρισμό, το τέταρτο και το ένατο θέμα περιλαμβάνει λέξεις όπως: νησί, Σαντορίνη, καλοκαίρι, μετανάστες, χώρα, κόμμα, κοινωνία, ευρώ, κέντρο, Τσίπρας.



Σχήμα 5.7 Απεικόνιση του δέκατου θέματος της τρίτης ομάδας

Συμπεράσματα και μελλοντική έρευνα

Συμπεράσματα

Σε αυτή την διπλωματική μελέτη αναλύθηκε η σημαντικότητα των δεδομένων του δημοφιλέστερου ρωσικού κοινωνικού δικτύου. Το ενδιαφέρον των ρωσόφωνων χρηστών για την Ελλάδα αναδείχθηκε μέσω των ποικίλων θεμάτων που εντοπίστηκαν στις αναρτήσεις του «VK». Στην συνέχεια περιεγράφηκε λεπτομερώς η διαδικασία μέσω της οποίας συλλέχθηκαν τα απαιτούμενα δεδομένα.

Για την επεξεργασία των αναρτήσεων χρησιμοποιήθηκαν κατάλληλοι αλγόριθμοι για τον μετασχηματισμό τους σε εμφυτεύσεις με τέτοιο τρόπο, ώστε να διατηρηθεί η σημασιολογική τους ταυτότητα. Σαν αποτέλεσμα λήφθηκε μοντέλο το οποίο επαληθεύθηκε για την εγκυρότητα των αποτελεσμάτων, με αποτέλεσμα να είναι εφικτή η μετέπειτα ανάλυση.

Τα διανύσματα που αντιπροσωπεύουν τις λέξεις ομαδοποιήθηκαν με τον K-Means αλγόριθμο και για λόγους οπτικοποίησης των 300 διαστάσεων, εφαρμόστηκε η ανάλυση κύριων συνιστωσών. Τέλος, αναλύθηκαν τα κύρια θέματα που εμπεριέχονται στην κάθε ομάδα. Έτσι, ο συνδυασμός των διαδικασιών που περιεγράφηκαν προσέφερε μια πλήρη εικόνα για τα αρχικά δεδομένα.

Μελλοντικές επεκτάσεις

Με την χρήση περισσότερου όγκου δεδομένων το Doc2vec μοντέλο θα μπορούσε να γίνει πιο ακριβές και ευέλικτο, με αποτέλεσμα η τελική ανάλυση να είναι η βέλτιστη δυνατή. Ενώ ένας πιο λεπτομερής «καθαρισμός» των λέξεων που δεν φέρνουν σημαντικό πληροφοριακό περιεχόμενο θα μπορούσε να φέρει καλύτερα αποτελέσματα.

Επειδή το «VK» με τον καιρό αποκτάει όλο και περισσότερη δημοτικότητα, δημιουργούνται καινούργιες σελίδες, και από τους απλούς χρήστες, αλλά και από τις εταιρείες, τα καινούργια δεδομένα πιθανόν να φέρουν σημαντικότερες και πιο χρήσιμες πληροφορίες.

Βιβλιογραφία

[1] Alexa (2017), https://www.alexa.com/siteinfo/vk.com

[2] Arthur, D.; Vassilvitskii, S. (2007). "k-means++: the advantages of careful seeding" (PDF). Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035

[3] BaseGroup Labs - Технологии анализа данных. https://basegroup.ru/community/glossary/k-means

[4] C.Sievert, K.E.Shirley (2014). *LDAvis: A method for visualizing and interpreting topics*

[5] C. Moody (2016). *Word2vec, LDA, and introducing a new hybrid algorithm: lda2vec,* https://www.slideshare.net/ChristopherMoody3/word2vec-lda-and-introducing-a-new-hybrid-algorithm-lda2vec

[6] CNN Greece. (2017). *Αυξήθηκαν οι Ρώσοι τουρίστες προς την Ελλάδα,* http://www.cnn.gr/taksidi/story/70908/ayxithikan-oi-rosoi-toyristes-pros-tin-ellada

[7] D.Dus (2016). *word2vec*, https://www.slideshare.net/DenisDus1/word2vec-part-1-61489929

[8] ENTERPRICE GREECE (2014). *Τουρισμός*, http://www.enterprisegreece.gov.gr/gr/ependyseis/ependytikoi-tomeis/toyrismos

[9] iefimerida. (2016). *Σχεδόν 1 εκατομμύριο Ρώσοι τουρίστες επισκέφθηκαν την Ελλάδα το 2016,* http://www.iefimerida.gr/news/307169/shedon-1-ekatommyrio-rosoi-toyristes-episkefthikan-tin-ellada-2016#axzz4gxO6HrU8

[10] J.Chuang, C.D.Manning, J.Heer (2012). *Termite: Visualization Techniques for Assessing Textual Topic Models*

[11] L.Konstantinovsky (2016). *Case Study: Text Processing with gensim through word2vec and doc2vec algorithms*, https://events.yandex.ru/lib/talks/4137/

[12] Lenta.ru. (2012). https://lenta.ru/lib/14204723/

[13] M.Bonzanini (2017). *Word Embeddings for Natural Language Processing in Python @Data Science Festival 2017*,

https://speakerdeck.com/marcobonzanini/word-embeddings-for-natural-language-processing-in-python-at-data-science-festival-2017

[14] NSS (2017). *An Intuitive Understanding of Word Embeddings: From Count Vectors to Word2Vec,* https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

[15] Quoc Le, T.Mikolov (2014). *Distributed Representations of Sentences and Documents*

[16] R.Řehůřek. (2009). https://radimrehurek.com/gensim/

[17] S. Nikolenko (2012), https://habrahabr.ru/company/surfingbird/blog/150607/

[18] StatSoft, *Кластеризация: метод k-средних,* http://statistica.ru/theory/klasterizatsiya-metod-k-srednikh/

[19] T.Mikolov (2016). *Case Study: Distributed Representation for NLP*, https://www.slideshare.net/mlprague/tom-mikolov-distributed-representations-for-nlp

[20] VCIOM. (2016). Έκδοση Τύπου № 3084, https://wciom.ru/index.php?id=236&uid=115657

[21] Vkontakte : https://vk.com/dev/methods

[22] Γ.Απλαδάς. (2014). Διαχείριση Δεδομένων & Κοινωνικά Δίκτυα στον Τουρισμό, https://ma.ellak.gr/documents/2014/05/%CE%B4%CE%B9%CE%B1%CF%87%CE%B5%CE%AF%CF%81%CE%B9%CF%83%CE%B7-%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CF%89%CE%BD-%CE%BA%CE%BF%CE%B9%CE%BD%CF%89%CE%BD%CE%B9%CE%BA%CE%AC-%CE%B4%CE%AF%CE%BA%CF%84.pdf

[23] Η ΚΑΘΗΜΕΡΙΝΗ. (2017). *Η αεπιλή της τρομοκρατίας σε Αίγυπτο, Τουρκία ευνοεί τον ελληνικό τουρισμό,* http://www.kathimerini.gr/860651/article/epikairothta/kosmos/h-apeilh-ths-tromokratias-se-aigypto-toyrkia-eynoei-ton-ellhniko-toyrismo

[24] А.С. Коляда, В.А Яковенко, В.Д. Гогунский (2014). *ПРИМЕНЕНИЕ ЛАТЕНТНОГО РАЗМЕЩЕНИЯ ДИРИХЛЕ ДЛЯ АНАЛИЗА ПУБЛИКАЦИЙ ИЗ НАУКОМЕТРИЧЕСКИХ БАЗ ДАННЫХ, УДК 004.62*

[25] Бурко Р.А., Терёшина Т.В. (2014). *Социальные сети в современном обществе // Молодой ученый.*. — №7. 607-608.

[26] Национальный Открытый Университет "ИНТУИТ". *Лекция 86: Латентное размещение Дирихле,*
http://www.intuit.ru/studies/courses/13844/1241/lecture/27038

[27] С.Николенко (2014). *Вероятностные модели: от наивного Байеса к LDA*, https://habrahabr.ru/company/surfingbird/blog/228249/

[28] С.Николенко (2014). *Категоризация текстов и модель LDA*,
https://logic.pdmi.ras.ru/~sergey/teaching/mlkfu14/17-lda.pdf

[29] Салмин А.А. (2013). *Анализ данных. Конспект лекций.* ФГОБУ ВПО «ПГУТИ»