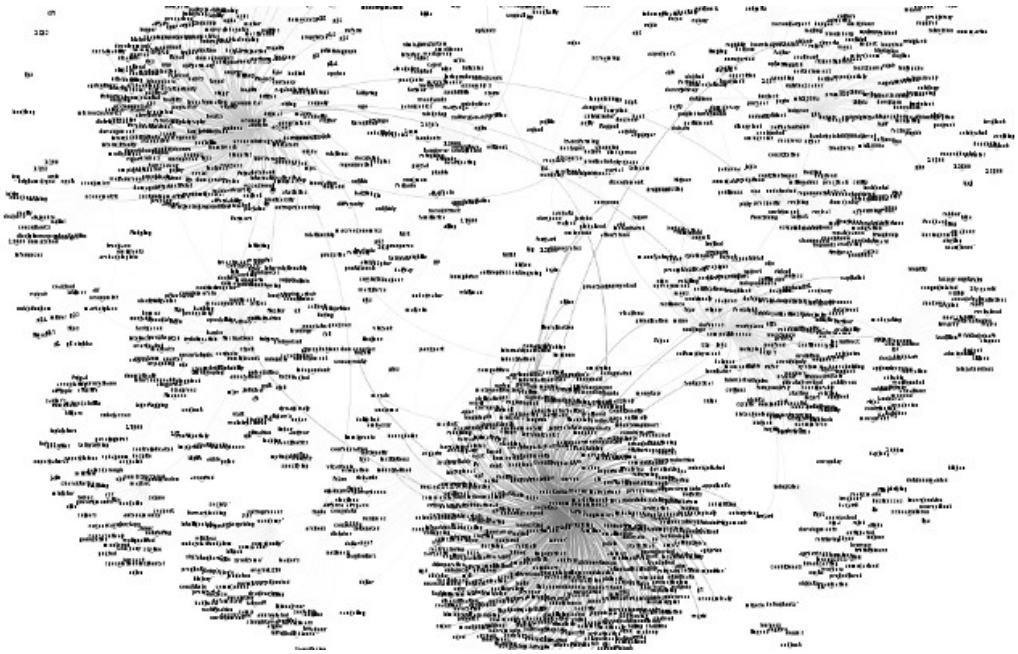
Αναγνώριση Συγγραφέα με χρήση Κατανεμημένων Αναπαραστάσεων Κειμένων (Authorship Attribution Using Distributed Document Representations)



Η Διπλωματική Εργασία

παρουσιάστηκε ενώπιον

του Διδακτικού Προσωπικού του

Πανεπιστημίου Αιγαίου

σε μερική εκπλήρωση των απαιτήσεων για το μεταπτυχιακό δίπλωμα τεχνολογίες και διοίκηση πληροφοριακών και επικοινωνιακών συστημάτων

του

Κλεάρχου Κτίστου

Η ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ ΔΙΔΑΣΚΟΝΤΩΝ ΕΠΙΚΥΡΩΝΕΙ

ΤΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΤΟΥ ΚΛΕΑΡΧΟΥ ΚΤΙΣΤΟΥ

Ευστάθιος Σταματάτος, Επιβλέπων

Τμήμα Μηχανικών Πληροφοριακών και

Επικοινωνιακών Συστημάτων

Χρήστος Γκουμόπουλος, Μέλος

Τμήμα Μηχανικών Πληροφοριακών και

Επικοινωνιακών Συστημάτων

Ανδρέας Παπασαλούρος, Μέλος

Τμήμα Μαθηματικών

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Πίνακας περιεχομένων

Πίνακας περιεχομένων 3
Περίληψη 5
Abstract 6
Κεφάλαιο 1 7
Εισαγωγικά στοιχεία 7
1.1 Εισαγωγή 7
1.2 Ανάπτυξη της τεχνικής 7
1.3 Κατανεμημένη αναπαράσταση 9
1.4 Σχετικές εργασίες 10
Κεφάλαιο 2 21
Ανάθεση συγγραφέα χρησιμοποιώντας κατανεμημένη αναπαράσταση εγγράφων . 21
2.1 Κατανεμημένη αναπαράσταση των εγγράφων 21
2.2 Απόδοση συγγραφέα ως πρόβλημα ταξινόμησης 23
Κεφάλαιο 3 25
Το Doc2vec 25
3.1 Εισαγωγή στο Doc2Vec 25
3.2 word2vec 26
3.3 Αλγόριθμοι Word2vec 26
3.3.1 CBOW 26
3.3.2 Το Skip gram 27
3.4 Doc2vec 28
3.5 Αξιολόγηση 30
Κεφάλαιο 4 31
Χρήση του Doc2Vec 31
4.1 Περίληψη 31
4.2 Ανάλυση 32
4.2.1 Εισαγωγή 33
4.2.2 Εκπαίδευση 34
4.2.3 Χρήση μνήμης 34
4.2.4 I / O 35
4.3 Θεματικά μοντέλα 38
4.3.1 Πλεονεκτήματα 39
4.3.2 Μειονεκτήματα 39
4.4 Διανύσματα παραγράφου (doc2vec) 39
4.4.1 Πλεονεκτήματα 41
4.4.2 Μειονεκτήματα 42
4.5 Διανύσματα Skip-Thought 42

4.5.1 Πλεονεκτήματα 43
4.5.2 Μειονεκτήματα 43
4.6 FastSent 43
4.6.1 Πλεονεκτήματα 44
4.6.2 Μειονεκτήματα 44
4.7 Σταθμισμένο άθροισμα των διανυσμάτων λέξεων 44
4.7.1 Πλεονεκτήματα 45
4.7.2 Μειονεκτήματα 45
Κεφάλαιο 5 46
Αλγόριθμος 46
Παρουσίαση κώδικα 46
Κεφάλαιο 6 52
Αποτελέσματα 52
Κεφάλαιο 7 56
Συμπεράσματα 56
Εικόνες/ Πίνακες 57
Βιβλιογραφία 58

Περίληψη

Η κατανεμημένη αναπαράσταση λέξεων σε ένα χώρο διανυσμάτων είναι μια τεχνική που επιτρέπει να αναπαρίστανται λέξεις με τη μέθοδο της γειτνίασης. Οι κατανεμημένες αναπαραστάσεις μπορούν να επεκταθούν σε μεγαλύτερες δομές όπως φράσεις, προτάσεις, έγγραφα. Η ικανότητα κωδικοποίησης πληροφοριών κειμένου και η δυνατότητα χειρισμού δεδομένων μεγάλης διάστασης είναι οι λόγοι για τους οποίους ο τρόπος αυτός χρησιμοποιείται ευρέως σε διάφορες εργασίες επεξεργασίας φυσικών γλωσσών,

Σε αυτή την εργασία, θα προσπαθήσουμε να κάνουμε την χρήση κατανεμημένης αναπαράστασης σε έγγραφα με σκοπό την εύρεση – ταυτοποίηση του συντάκτη.. Η προτεινόμενη μέθοδος χρησιμοποιεί κατανεμημένες αναπαραστάσεις διανυσμάτων σε έγγραφα και στη συνέχεια χρησιμοποιεί n-grams για να εκτελέσει την αυτόματη απόδοση σε συγγραφέα.

Θα χρησιμοποιήσουμε n-grams ως δεδομένα εισόδου για το μοντέλο. Θα πραγματοποιηθεί έλεγχος σε κείμενα με σκοπό να επαληθεύομε το μοντέλο ή ακόμα και να βγάλουμε καλυτέρα αποτελέσματα.

Λέξεις κλειδιά: n-grams, doc2vec,word recognition

Abstract

Distributed word representation in a vector space is a technique that allows words to be represented with a proximity method. Distributed representations can be extended to larger structures such as phrases, sentences, documents. The ability to encode text information and the ability to manipulate large-dimensional data are the reasons why this mode is widely used in various language-editing tasks,

In this thesis, we will try to make use of a distributed representation in documents to find and identify the author. The proposed method uses distributed vector representations in documents and then uses n-grams to perform automatic writer performance.

We will use n-grams as input data for the model. We do not carry out a control in a wagon in order to impregnate the model or even get better results.

KEY WORDS: n-grams, doc2vec,word recognition

Κεφάλαιο 1

Εισαγωγικά στοιχεία

1.1 Εισαγωγή

Η ενσωμάτωση λέξεων είναι το όνομα μιας σειράς γλωσσικών μοντέλων και χαρακτηριστικών τεχνικών μάθησης στη επεξεργασία φυσικής γλώσσας (NLP) όπου λέξεις ή φράσεις από το λεξιλόγιο χαρτογραφούνται σε φορείς πραγματικών αριθμών. Εννοείται ότι περιλαμβάνει μια μαθηματική ενσωμάτωση από ένα χώρο με μία διάσταση ανά λέξη σε ένα συνεχές χώρο με πολύ μικρότερη διάσταση.

Οι μέθοδοι για τη δημιουργία αυτής της χαρτογράφησης περιλαμβάνουν τα νευρωνικά δίκτυα, [1] τη μείωση της διαστασιολόγησης στη μήτρα συσχέτισης λέξεων, [2] [3] [4] μοντέλα πιθανοτήτων, [5] εξηγηθείσα μέθοδο βάσης γνώσεων, [6] του πλαισίου στο οποίο εμφανίζονται τα λόγια [7].

Οι ενσωματώσεις λέξεων και φράσεων, όταν χρησιμοποιούνται ως η υποκείμενη αναπαράσταση εισόδου, έχουν αποδειχθεί ότι ενισχύουν την απόδοση σε εργασίες NLP(Natural Language Process), όπως η συντακτική ανάλυση [8] και η ανάλυση συναισθημάτων. [9]

1.2 Ανάπτυξη της τεχνικής

Στη γλωσσολογία συζητήθηκαν λέξεις ενσωματωμένες στον ερευνητικό τομέα της σημασιολογικής κατανομής. Σκοπός του είναι να ποσοτικοποιήσει και να κατηγοριοποιήσει τις σημασιολογικές ομοιότητες μεταξύ γλωσσικών αντικειμένων με βάση τις ιδιότητες κατανομής τους σε μεγάλα δείγματα γλωσσικών δεδομένων. Η υποκείμενη ιδέα ότι "μια λέξη χαρακτηρίζεται από την βάση που κρατάει" ήταν γνωστή από τον Firth. [10]

Η τεχνική της έκφρασης των λέξεων ως φορέων έχει τις ρίζες της στη δεκαετία του 1960 με την ανάπτυξη του μοντέλου διανυσματικού χώρου για την ανάκτηση πληροφοριών. Η μείωση του αριθμού των διαστάσεων με την αποσύνθεση της μοναδικής τιμής οδήγησε στην εισαγωγή της λανθάνουσας σημασιολογικής ανάλυσης

στα τέλη της δεκαετίας του 1980[1]. Το 2000, οι Bengio et al. παρείχε σε μια σειρά εγγράφων τα "Νευρωνικά μοντέλα πιθανοτήτων γλώσσας" για να μειώσουν την υψηλή διαστασιολόγηση των εκφράσεων/λέξεων σε περιβάλλοντα με την "εκμάθηση κατανεμημένης αναπαράστασης λέξεων". (Bengio et al, 2003). [11] Οι ενσωματώσεις λέξεων έρχονται σε δύο διαφορετικές μορφές, μία στην οποία τα λόγια εκφράζονται ως φορείς διαδοχικών λέξεων και μία άλλη στην οποία οι λέξεις εκφράζονται ως φορείς των γλωσσικών πλαισίων στα οποία εμφανίζονται τα λόγια. Αυτά τα διαφορετικά στυλ μελετώνται στο (Lavelli et al, 2004). [12] Οι Roweis και Saul δημοσίευσαν στο περιοδικό "Science" πώς χρησιμοποιώντας την "τοπική γραμμική ενσωμάτωση" (LLE) ανακάλυψαν παραστάσεις δομών δεδομένων μεγάλης διάστασης. [13] Η περιοχή αναπτύχθηκε σταδιακά και πραγματικά απογειώθηκε μετά το 2010, εν μέρει λόγω της σημαντικής προόδου που σημειώθηκε από τότε στην ποιότητα των φορέων και την ταχύτητα εκπαίδευσης του μοντέλου.

Υπάρχουν πολλοί κλάδοι και πολλές ερευνητικές ομάδες που εργάζονται στην ενσωμάτωση λέξεων. Το 2013, μια ομάδα της Google με επικεφαλή τον Tomas Mikolov δημιούργησε το word2vec, ένα εργαλείο ενσωμάτωσης λέξεων που μπορεί να εκπαιδεύσει τα μοντέλα διανυσματικών χώρων γρηγορότερα από τις προηγούμενες προσεγγίσεις. Οι περισσότερες νέες τεχνικές ενσωμάτωσης λέξεων βασίζονται σε αρχιτεκτονική νευρωνικών δικτύων αντί σε πιο παραδοσιακά μοντέλα n-gram και σε μη επιτηρούμενη μάθηση[2].

Ένας από τους κύριους περιορισμούς των ενσωματώσεων λέξεων (μοντέλα διαστήματος, διάνυσμα λέξεων εν γένει) είναι ότι οι πιθανές σημασίες μιας λέξης συγχωνεύονται σε μία μόνο αναπαράσταση (έναν μόνο φορέα στον σημασιολογικό χώρο). Οι αισθητηριακές ενσωματώσεις[3] είναι μια λύση στο πρόβλημα αυτό: οι μεμονωμένες έννοιες των λέξεων αντιπροσωπεύονται ως διακριτοί φορείς στον χώρο.

[1] **Sahlgren Magnus. "A brief history of word embeddings".**

[2] **doc2vec vector https://medium.com/scaleabout/a-gentle-introduction-to-doc2vec-db3e8c0cce5e**

[3] **Overview of embedding methods http://mlexplained.com/2017/12/28/an-overview-of-sentence-embedding-methods/**

1.3 Κατανεμημένη αναπαράσταση

Η κατανεμημένη αναπαράσταση λέξεων σε ένα διάνυσμα χώρου, επίσης γνωστή ως ενσωμάτωση λέξεων (word embeddings) (Turian et al., 2010, Pennington et al., 2014), είναι μία νέα τεχνική που σήμερα χρησιμοποιείται ευρέως σε πολλές εργασίες επεξεργασίας φυσικής γλώσσας (NLP). Σκοπός της είναι να αντιπροσωπεύει λέξεις με σταθερό μήκος, σε συνεχή και πυκνά διανύσματα χαρακτηριστικών.

Ένα πολύ δημοφιλές μοντέλο αρχιτεκτονικής εκμάθησης κατανεμημένων αναπαραστάσεων διανυσμάτων λέξεων (Word2Vec) χρησιμοποιώντας ένα νευρωνικό δίκτυο προτάθηκε από τον Mikolov et al. (2013α, β). Αυτή η τεχνική συλλαμβάνει τη σημασιολογία και τις συντακτικές σχέσεις των λέξεων: δηλ παρόμοιες λέξεις, όταν είναι κοντά η μία στην άλλη στον διανυσματικό χώρο. Παραδείγματος χάριν, παρουσιάστηκε από τον Mikolov et al. (2013c) ότι το διάνυσμα [King] - διάνυσμα [Man] + διάνυσμα [Woman] έχει ως αποτέλεσμα στον διανυσματικό χώρο να είναι πιο κοντά στην αναπαράσταση του διανύσματος [Queen].

Άλλες δύο γνωστές συνεχείς αναπαραστάσεις λέξεων είναι η λανθάνουσα σημασιολογική ανάλυση (LSA) (Wiemer-Hastings et al., 2004) και η λανθάνουσα κατανομή Dirichlet (LDA) (Trejo et al., 2015).Σε αντίθεση με LSA και LDA, οι αναπαραστάσεις διανυσμάτων που αποκτήθηκαν από τον Mikolov et al. (2013α, β) διατηρούν τις γραμμικές κανονικότητες μεταξύ των λέξεων. Η επιτυχία αυτής της τεχνικής εξηγείται από: (1) την ικανότητά της να κωδικοποιεί την σημασιολογική πληροφορία των λέξεων σε φορείς, που την καθιστά κατάλληλη για ανάλυση συναισθημάτων (Socher et al., 2013a), συντακτική ανάλυση (Socher et al. et al., 2014) συνοπτική παρουσίαση (Miranda et al., 2014) και πολλές άλλες εφαρμογές στον τομέα της επεξεργασίας φυσικής γλώσσας. και (2) στην ικανότητά του να χειρίζεται σύνολα δεδομένων μεγάλης διάστασης (από χιλιάδες έως εκατομμύρια περιπτώσεις).

Οι κατανεμημένες αναπαραστάσεις μπορούν να επεκταθούν σε μοντέλα και όχι μόνο σε λέξεις, αλλά και σε μεγαλύτερες δομές γλώσσας όπως φράσεις, προτάσεις και έγγραφα (Le και Mikolov 2014).

Σε αυτή την εργασία, θα ασχοληθούμε με το Doc2vec (Le και Mikolov 2014), η οποία δημιουργεί μια κατανομή διανυσματικής αναπαράστασης σε επίπεδο εγγράφου

Θα χρησιμοποιήσουμε n-grams ως χαρακτηριστικά για την κατασκευή της αναπαράστασης κατανομής κειμένου αντί να χρησιμοποιούμε μόνο λέξεις, έτσι ώστε το μοντέλο να μάθει τα συντακτικά και τα γραμματικά πρότυπα ενός συγγραφέα.

Γενικά, θεωρείται ότι για την απόκτηση ακριβών κατανεμημένων αναπαραστάσεων, απαιτείται ένα πολύ μεγάλο σώμα για εκπαίδευση Εμείς στην εργασία μας πειραματιστήκαμε με λίγα μόνο κείμενα.

Η χρήση της αναπαράστασης κατανομής κειμένου σπανίως εφαρμόζεται στην εφαρμογή ανάθεσης συγγραφέων και τα προηγούμενα αποτελέσματα που χρησιμοποιούν αυτή την τεχνική είναι αποθαρρυντικά (Ρόδος 2015). Αξίζει να σημειωθεί ότι σε γενικές γραμμές αυτή η τεχνική σπάνια χρησιμοποιείται σε εργασίες επεξεργασίας φυσικής γλώσσας με μικρά κείμενα.

1.4 Σχετικές εργασίες

Η μέθοδος απόδοσης συντάκτη, γνωστή και ως εφαρμογή αναγνώρισης του συντάκτη, συνίσταται στην αναγνώριση του δημιουργού ενός δεδομένου κειμένου. Μπορεί να θεωρηθεί ως πρόβλημα ταξινόμησης, όταν δοθεί ένα σύνολο εγγράφων που ανήκουν σε διάφορους συντάκτες και ένα σύνολο εγγράφων με άγνωστους συγγραφείς. Είναι απαραίτητο να καθοριστούν οι αντίστοιχοι συγγραφείς των εγγράφων με τα άγνωστου συγγραφέα κείμενα, με άλλα λόγια, να επιλέξουν την τάξη (τον συντάκτη) που αντιστοιχεί σε κάθε άγνωστο έγγραφο. Η διαδικασία απόδοσής του συντάκτη είναι ένας ενεργός ερευνητικός τομέας με πολλές σχετικές εφαρμογές όπως ανίχνευση λογοκλοπής (Stamatatos 2011, Sanchez-Perez et al., 2015), ταυτοποίηση συγγραφέων, για να κατονομάσουμε ορισμένα π.χ. σε σημειώματα αυτοκτονίας, σημειώματα λύτρων, απειλητικά μηνύματα ηλεκτρονικού ταχυδρομείου,

(Gómez Adorno et 2016, Chaski 2005, Brocardo et al., 2013) και το προφίλ του συγγραφέα (Argamonet αϊ. 2009).

Προκειμένου να επιτευχθεί η ανάθεση συγγραφέων, είναι απαραίτητο να προσδιοριστούν τα χαρακτηριστικά ή τα προφίλ που αντιστοιχούν στους συγκεκριμένους δημιουργούς. Ο χαρακτηρισμός του στυλ γραφής ενός συγγραφέα παραμένει ένα ανοιχτό πρόβλημα, το οποίο μελετάται από την ερευνητική αρχή και ονομάζεται στυλομετρία. . Τα χαρακτηριστικά που αποτυπώνουν το στυλ ενός συγγραφέα είναι γνωστά ως δείκτες στυλ και υποθέτουμε ότι είναι αρκετά ισχυροί για να μοντελοποιήσουν το στυλ κάτω από διαφορετικές συνθήκες, δηλαδή το στυλ διατηρείται σε διαφορετικά είδη και σε βάθος χρόνου (Alzahrani et al., 2012).

Οι δείκτες στυλ μπορούν να ταξινομηθούν ανάλογα με τις πληροφορίες που χρησιμοποιούν. δηλαδή, μπορούν να αντιστοιχηθούν στα ακόλουθα επίπεδα: χαρακτήρα, λεξικό, συντακτικό, σημασιολογικό και μορφοποίηση. Οι δείκτες στυλ στο χαρακτήρα και τα λεξικά επίπεδα δεν χρειάζονται σύνθετη επεξεργασία για να αποκτηθούν και παρουσιάζουν καλά αποτελέσματα. Άλλοι δείκτες στυλ απαιτούν πολύπλοκη επεξεργασία και τα αποτελέσματά τους θεωρούνται συνήθως χαμηλότερα (Stamatatos 2009), αλλά σε πρόσφατη εργασία (Sidorov et al., 2014, Posadas- Duran et αϊ. 2014; Posadas-Durán et αϊ. 2015), η χρήση συντακτικών n-grams που εκμεταλλεύονται τις πληροφορίες των δέντρων εξάρτησης είχε υψηλότερα αποτελέσματα από τα n-grams ετικέτας.

Ένα από τα παλαιότερα μοντέλα, το τυπικό μοντέλο bag-of -words (Stamatatos 2009), συνίσταται στην εκπροσώπηση ενός κειμένου σαν σύνολο λέξεων με τις αντίστοιχες συχνότητες τους, δηλ. Οι λέξεις είναι χαρακτηριστικά στο αντίστοιχο μοντέλο διανυσματικού χώρου. Σε αυτό το μοντέλο, θεωρείται επίσης ότι η εμφάνισή τους είναι ανεξάρτητη η μία από την άλλη. Για το πρόβλημα της ανάθεσης του συντάκτη, οι πιο συχνές λέξεις (λέξεις λειτουργίας ή λέξεις διακοπής) αποδείχτηκαν να έχουν μεγάλη ακρίβεια.

Από την άλλη πλευρά, τα χαρακτηριστικά που βασίζονται σε λέξεις n-grams προτάθηκαν για να επωφεληθούν από τις πληροφορίες συμφραζομένων και τα αποτελέσματα που έχουν αποκτηθεί είναι ελαφρώς καλύτερα από του μοντέλου bag-of-words.

Ο πιο ακριβής δείκτης στυλ για την απόδοση του συντάκτη είναι οι χαρακτήρες n-grams (Stamatatos et al., 2001, Kešelj et al., 2003, Sidorov et al., 2014). Αυτοί οι δείκτες στυλ δείχνουν υψηλή ακρίβεια για διάφορα συγκριτικά σημεία (Stamatatos 2013, Plakias και Stamatatos 2008, Escalante et al., 2011). Ωστόσο, το μειονέκτημα τους είναι ότι δημιουργούν πολύ αραιά διανύσματα μεγάλων διαστάσεων. Εκτός αυτού, δημιουργούν αναπαραστάσεις κειμένου χωρίς σαφή σημασιολογική ερμηνεία.

Ο Stamatatos (2009), αναφέρει ότι η χρήση σημασιολογικών χαρακτηριστικών για την απόδοση συγγραφέα συνήθως βελτιώνει τα ληφθέντα αποτελέσματα. Ωστόσο, πολύ λίγες προσπάθειες έχουν γίνει για την αξιοποίηση χαρακτηριστικών σε υψηλό επίπεδο για λόγους μεθοδολογίας. Σε αυτή την εργασία, εξετάζουμε τη χρήση της αναπαράστασης κατανεμημένου εγγράφου για το απόδοση συγγραφέων, το οποίο αντιστοιχεί επακριβώς στα σημασιολογικά χαρακτηριστικά.

Στην εργασία (Segarra et al., 2013), οι συγγραφείς προτείνουν δίκτυα γειτονίας λέξεων (WANs), όπου οι κόμβοι είναι λέξεις-κλειδιά και οι κατευθυνόμενες ακμές αντιπροσωπεύουν την πιθανότητα εύρεσης λέξης λέξη-στόχο στη διαταγμένη εγγύτητα μιας λέξης λειτουργίας. Στην εργασία αυτή, οι συγγραφείς αναφέρουν ότι η ακρίβεια που επιτυγχάνεται από τα δίκτυα WAN είναι υψηλότερη από αυτή που επιτυγχάνεται με τις παραδοσιακές μεθοδολογίες που βασίζονται μόνο στις συχνότητες των λέξεων λειτουργίας. Παρόλο που τα δίκτυα WAN επέτυχαν υψηλή ακρίβεια σε πολύ μακρά κείμενα όσο μια θεατρική πράξη ή ένα μυθιστόρημα, λαμβάνουν μόνο φυσιολογικές τιμές για σύντομα κείμενα, όπως τα τεύχη μιας εφημερίδας, εάν ο αριθμός των υποψηφίων συγγραφέων είναι μικρός.

Πρόσφατα, η χρήση κατανεμημένης αναπαράστασης έχει δείξει μεγάλη δύναμη στη συλλογή της σημασιολογίας λέξεων, φράσεων και προτάσεων, που ωφελούν τις εφαρμογές επεξεργασίας φυσικής γλώσσας. Οι Le και Mikolov (2014), παρουσιάζουν το Doc2vec, έναν αλγόριθμο χωρίς επίβλεψη που μαθαίνει αναπαραστάσεις χαρακτηριστικών σταθερού μήκους από έγγραφα μεταβλητού μήκους. Η ιδέα είναι να συνδυαστεί η έννοια των λέξεων για την κατασκευή της έννοιας των εγγράφων χρησιμοποιώντας το μοντέλο κατανεμημένης μνήμης. Η κατανεμημένη αναπαράσταση που επιτυγχάνεται με τη μέθοδο Doc2vec υπερέχει τόσο των μοντέλων bag-of-words όσο και των λέξεων n-grams παράγοντας νέα αποτελέσματα για πολλές εργασίες ταξινόμησης κειμένων και συναισθημάτων.

Σε μια άλλη σχετική εργασία (Li και Shindo 2015), οι συγγραφείς παρουσιάζουν μια ελεγχόμενη μέθοδο που ονομάζεται Compound RNN. Χρησιμοποιεί επαναληπτικά και επαναλαμβανόμενα νευρωνικά δίκτυα για να μάθει την κατανεμημένη αναπαράσταση του εγγράφου. Αυτή η μέθοδος είναι συγκεκριμένης εργασίας επειδή δεν μαθαίνει τη γενική αναπαράσταση των προτάσεων όπως στην περίπτωση του doc2vec. Παρ 'όλα αυτά, αυτή η μέθοδος ξεπερνά τις υπάρχουσες βασικές γραμμές σε εργασίες όπως η δυαδική ταξινόμηση, η ταξινόμηση πολλαπλών τάξεων και η παλινδρόμηση[16].

Μια άλλη ενδιαφέρουσα εφαρμογή που χρησιμοποιεί το μοντέλο Doc2vec και το skip-gram για την διάκριση παρόμοιων γλωσσών παρουσιάζεται στους Franco-Salvador et al. (2015). Οι συγγραφείς χρησιμοποιούν το μοντέλο συνεχών skip-gram (Word2vec) (Mikolov et al., 2013b) για να παράγουν κατανεμημένες αναπαραστάσεις λέξεων (δηλ. φορείς N διαστάσεων) και να υπολογίζουν το μέσο όρο των διαστάσεών τους για να δημιουργήσουν την κατανεμημένη αναπαράσταση εγγράφων. Αξιολογούν επίσης το μοντέλο ταξινόμησης τους χρησιμοποιώντας τη μέθοδο Doc2vec για την εκμάθηση της αναπαράστασης εγγράφων. Για το σκοπό αυτό, ο συνδυασμός των διανυσμάτων λέξεων (Word2vec) έλαβε καλύτερα αποτελέσματα κατά μέσο όρο σε σύγκριση με τη χρήση των διανυσμάτων που δημιουργήθηκαν απευθείας από προτάσεις (Doc2vec). Παρόλα αυτά, αυτό το συμπέρασμα ισχύει για τη διαδικασία αναγνώρισης γλωσσικών παραλλαγών, αλλά για την απόδοση του συντάκτη, αυτή η προσέγγιση δεν είναι εφαρμόσιμη, όπως θα δείξουμε στις επόμενες ενότητες.

Υπάρχουν λίγες εφαρμογές χρήσης κατανεμημένων αναπαραστάσεων για την απόδοση του δημιουργού. Οι Kiros et al. (2014), προτείνουν ένα πλαίσιο για την εκμάθηση κατανεμημένων αναπαραστάσεων χαρακτηριστικών. Τα χαρακτηριστικά αντιστοιχίζονται σε μια ευρεία ποικιλία εννοιών, όπως είναι οι ενδείξεις εγγράφων (για να μάθουν τα διανύσματα των προτάσεων), οι γλωσσικοί δείκτες (για να μάθουν τις αντιπροσωπευτικές εκφράσεις γλώσσας), τα μεταδεδομένα και πρόσθετες πληροφορίες σχετικά με τους συγγραφείς (να μάθουν το προφίλ των συγγραφέων, όπως η ηλικία, το φύλο και το επάγγελμα). Το πλαίσιο αξιολογείται σε διάφορα καθήκοντα: την ταξινόμηση των συναισθημάτων, την ταξινόμηση των διαγλωσσικών εγγράφων και την απόδοση του συντάκτη σε blog. Για την ανάθεση συγγραφέων, η μεθοδολογία αξιολογείται σε ένα "σώμα" ιστολογιών και τα χαρακτηριστικά βασίζονται μόνο στα μεταδεδομένα του συντάκτη και όχι στα ίδια τα κείμενα.

Αντίθετα, στη δική μας δουλειά χρησιμοποιούμε μόνο τις πληροφορίες κειμένου για να εξαγάγουμε αυτόματα σημαντικές παραστάσεις διανυσμάτων εγγράφων[16].

Μια καινοτόμος αρχιτεκτονική νευρωνικού δικτύου, δηλαδή συνεκτικά νευρωνικά δίκτυα (CNN), πάνω από τις λέξεις ενσωμάτωσης ήταν που παρουσιάστηκε στη Ρόδο το 2015. Αξιολογήθηκε σε δύο σύνολα δεδομένων: μια βασική γραμμή που αναπτύχθηκε από τον συγγραφέα και το σύνολο δεδομένων PAN 2012 (Juola 2012). Αυτή η δουλειά είναι άμεσα συγκρίσιμη με τη δική μας. Για την αναπαράσταση των εγγράφων, ο συγγραφέας χρησιμοποίησε το σύνολο των διανυσμάτων Google Newsword που εκπαιδεύτηκαν μέσω της μεθόδου skipgram και της αρνητικής δειγματοληψίας που παρουσιάστηκε από τον Mikolov et al. (2013α, β).

Ο συγγραφέας χρησιμοποίησε την τυπική προσέγγιση για τα συνεκτικά μοντέλα (απλή λειτουργία συγκαλλιέργειας) για να κωδικοποιήσει ακολουθίες παρά λέξεις, αντί να υπολογίζει κατά μέσο όρο τις διαστάσεις όπως έγινε στον Franco-Salvador et al. (2015). Η ταξινόμηση γίνεται μέσω λογιστικής παλινδρόμησης και τα αποτελέσματα δείχνουν υψηλή ακρίβεια σε σχέση με το σύνολο των δεδομένων της βάσης, αλλά η αρχιτεκτονική CNN δεν υπερέβη την καλύτερη μέθοδο που παρουσιάστηκε στο PAN 2012 (Juola 2012).

Κατά τη δημιουργία του αλγορίθμου μας προσπαθήσαμε μερικώς να ακολουθήσουμε τον τρόπο που ακολούθησαν και οι Posadas Duran et al στο paper με τίτλο **Application of the distributed document representation in the authorship attribution task for small corpora"**, όπου πραγματοποιούν πειράματα πάνω σε διαφορετικά σύνολα δεδομένων του corpus Guardian10 δημιουργώντας κατανεμημένες αναπαραστάσεις κειμένου με τη μέθοδο doc2vec των Le και Mikolov 2014, χρησιμοποιώντας μια προσέγγιση μηχανικής μάθησης.

Η εφαρμογή της μεθόδου Doc2vec απαιτεί τις ακόλουθες τρεις παραμέτρους: τον αριθμό των χαρακτηριστικών που πρέπει να επιστραφούν (μήκος του διανύσματος), το μέγεθος του παραθύρου που συλλαμβάνει τη γειτονιά και την ελάχιστη συχνότητα των λέξεων που πρέπει να ληφθούν υπόψη στο μοντέλο. Οι τιμές αυτών των παραμέτρων εξαρτώνται από το χρησιμοποιούμενο σώμα.

Η προτεινόμενη μέθοδος των Posadas Duran et al προέβλεπε για σκοπό της αύξησης της αποτελεσματικότητας της μεθόδου doc2vec εκπαίδευση του μοντέλου αρκετές φορές πάνω από το μη επισημασμένο σώμα, αλλά με εναλλαγή της σειράς των εισαγωγών στα έγγραφα. Πιο συγκεκριμένα στη εργασία τους αυτή προτάθηκαν εννέα διαφορετικές διαμορφώσεις, προκειμένου να διασφαλιστεί η αναπαραγωγικότητα των πειραμάτων, αντί να χρησιμοποιείτε μια γεννήτρια τυχαίων αριθμών. Οι κανόνες που ακολουθήθηκαν είναι οι εξής:

1. Πρώτος κανόνας: Ανατροπή της σειράς των στοιχείων στο σύνολο, δηλ. T = di, di-1,. . . , d1.

2. Δεύτερος κανόνας: Επιλογή πρώτα των εγγράφων με περιττό δείκτη εισόδου με αύξουσα σειρά και έπειτα έγγραφα με ζυγό δείκτη εισόδου, δηλ., T = [d1, d3,. . . , d2, d4,. . .].

3. Τρίτος κανόνας: Επιλογή πρώτα των εγγράφων με ζυγό δείκτη εισόδου σε αύξουσα σειρά και έπειτα τα έγγραφα με περιττό δείκτη, δηλ., T = [d2, d4,. . . , d1, d3,. . .].

4. Τέταρτος κανόνας: Για κάθε έγγραφο με μονό δείκτη i, γίνεται αντικατάσταση του με το έγγραφο με δείκτη i + 1, δηλαδή, T = [d2, d1, d4, d3,. . .].

5. Πέμπτος κανόνας: Μετακίνηση κάθε στοιχείου δύο θέσεις προς τα αριστερά με κυκλικό τρόπο, δηλ. Εάν m είναι ο δείκτης του τελευταίου στοιχείου του σώματος T τότε T = dm-1, dm, d1, d2,. . .

6. Έκτος κανόνας: Για κάθε έγγραφο με δείκτη i, ανταλλάζουμε με το έγγραφο του οποίου ο δείκτης είναι i + 3.

7. Έβδομος κανόνας: Για κάθε έγγραφο με δείκτη i του οποίου η τιμή είναι πολλαπλάσια των τριών, ανταλλάζουμε με το έγγραφο δίπλα του (i + 1).

8. Όγδοος κανόνας: Για κάθε έγγραφο με δείκτη i του οποίου η τιμή είναι πολλαπλάσια των τεσσάρων, ανταλλάζουμε το με το έγγραφο δίπλα σε αυτό (i + 1).

9. Ένατος κανόνας: Για κάθε έγγραφο με δείκτη i του οποίου η τιμή είναι πολλαπλάσια των τριών, ανταλλάζουμε με το έγγραφο του οποίου ο δείκτης είναι i + 2.

Τα πειράματα διεξήχθησαν χρησιμοποιώντας δύο διαφορετικούς ταξινομητές: μηχανές υποστήριξης διανυσμάτων (SVM) και γραμμική παλινδρόμηση (LR). Για τις περισσότερες περιπτώσεις, η γραμμική παλινδρόμηση απέδωσε τα καλύτερα αποτελέσματα. Επομένως, αναφέρονται μόνο τα αποτελέσματα που λαμβάνονται με αυτόν τον ταξινομητή.

Προκειμένου να δημιουργηθεί το μοντέλο κατανεμημένης αναπαράστασης, η μέθοδος Doc2vec λαμβάνει ως είσοδο τα έγγραφα απλού κειμένου του σετ εκπαίδευσης χωρίς καμία ετικέτα (δεν υπάρχει καμία περιγραφή της κλάσης στην οποία ανήκουν) και τα μη επισημασμένα έγγραφα κειμένου από το σετ δοκιμών.

Για προ επεξεργασία, πραγματοποιήθηκε μόνο μια τυπική διαδικασία tokenization. Διαφορετικές αναπαραστάσεις των κειμένων χρησιμοποιήθηκαν ως τύποι δεδομένων εισόδου για τη μέθοδο Doc2vec προκειμένου να αξιολογηθεί η ποιότητα των διαφορετικών εξόδων κατανεμημένης αναπαράστασης. Συγκεκριμένα, αναπαραστάθηκαν τα κείμενα με όρους λέξεις, 2-γραμμα, 3-γραμμα, 4-γραμμα και 5-γραμμα.

Για το πειραματικό σώμα του Guardian, χρησιμοποιήθηκε το δοκιμαστικό σχήμα που δημιουργήθηκε σε προηγούμενες εργασίες.

- Φάση κατάρτισης: Από τις πέντε διαφορετικές κατηγορίες του πειραματικού σώματος του Guardian (Πολιτική, Κοινωνία, Κόσμος, Βρετανία και κριτικές βιβλίων), η κατηγορία Πολιτική επιλέγεται ως εκπαιδευτικό σύνολο και το πολύ δέκα κείμενα ανά συγγραφέα χρησιμοποιούνται στην εκπαιδευτική φάση ταξινόμησης. Σύμφωνα με αυτό, ο συνολικός αριθμός των κειμένων που χρησιμοποιούνται στη φάση κατάρτισης (ntr) είναι 112.

- Φάση δοκιμών: Ο εκπαιδευμένος ταξινομητής δοκιμάζεται στις υπόλοιπες κατηγορίες, έχοντας συνολικά τέσσερα ζεύγη (Πολιτική-Κοινωνία, Πολιτική-Αγγλία, Πολιτική-Κόσμος, Πολιτική-Βιβλιοκρισίες) και επίσης και το πολύ 10 κείμενα ανά συγγραφέα επιλεγμένα.

Η ακρίβεια που επετεύχθη από τους ταξινομητές SVM και LR παρουσιάζεται στο παρακάτω σχήμα. Η απόδοση που επιτυγχάνεται από τον ταξινομητή LR είναι συγκρίσιμη με αυτή που αποκτήθηκε από τον ταξινομητή SVM, αλλά ο ταξινομητής LR πέτυχε υψηλότερη ακρίβεια. Στους ακόλουθους πίνακες που σχετίζονται με το πειραματικό σώμα The Guardian, αναφέρεται μόνο η ακρίβεια που έχει επιτευχθεί από τον ταξινομητή LR. Ο Πίνακας 1 παρουσιάζει τα αποτελέσματα που προκύπτουν για κάθε ζεύγος κατηγοριών μαζί με τον συνολικό αριθμό κειμένων που χρησιμοποιούνται σε κάθε κατηγορία και τα αποτελέσματα που προέκυψαν κατά την προηγούμενη εργασία (Char 3-grams Stamatatos 2013).

Name	Society	UK	World	Books	Average Acc.
D2V words	95.16	91.11	**93.16**	87.30	91.68
D2V words+2-grams	**98.38**	**95.55**	91.45	**93.65**	**94.75**
D2V words+2+3-grams	**98.38**	93.33	91,45	92.06	93.80
D2V words+2+3+4-grams	**98.38**	93.33	91,45	**93.65**	94.20
D2V words+2+3+4+5-grams	**98.38**	93.33	91,45	**93.65**	94.20
Char 3-grams (from Stamatatos 2013)	91.00	88.00	82,50	79.50	85.50
Με έντονη γραφή η υψηλότερη τιμή που επετεύχθη σε κάθε στήλη.					

Πίνακας 1. Αποτελέσματα του corpus Guardian υπολογίζοντας ως δεδομένα εκπαίδευσης των φάκελο politics και ως δεδομένα τεστ όλους τους υπόλοιπους φάκελους δεδομένων.

Στη συνέχεια περιγράφεται ένα διαφορετικό σχέδιο δοκιμής πάνω στο corpus The Guardian που προτείνεται από τους Sapkota et al. (2015). Εξετάζοντας το σύνολο όλων των συγγραφέων A = {A1, A2,. . . , Ai} και το σύνολο όλων των κατηγοριών C ={ CPolitics, CSsociety, CUK, CWorld, CBookReviews} όπου κάθε στοιχείο αντιπροσωπεύει όλα τα κείμενα σε μια κατηγορία, δηλαδή Clabel ={ t11,. . . tji με tji να είναι το i-th κείμενο του συγγραφέα Aj στην κατηγορία Clabel. Όπως και στο προηγουμένως εξειδικευμένο σχήμα, σε κάθε κατηγορία επιλέγονται το πολύ 10 κείμενα ανά συγγραφέα. Δεδομένου του συνόλου S που περιέχει όλα τα δυνατά 2-πολα που δημιουργούνται από τα στοιχεία του συνόλου C:

S ={(CPolitics, CSociety), (CPolitics, CUK), ...}

με πρωτεύοντα ίσο με 20, για κάθε sk ∈ S δημιουργείτε ένας ταξινομητή Fk χρησιμοποιώντας το πρώτο στοιχείο του 2-πολλου sk ως εκπαιδευτικού σετ και βρίσκεται η ακρίβεια του ταξινομητή Acck χρησιμοποιώντας το δεύτερο στοιχείο ως σύνολο δοκιμών. Τέλος, αναφέρεται η μέση ακρίβεια που επιτυγχάνεται από όλους τους ταξινομητές F.

Τα πειράματα πραγματοποιήθηκαν σε όλα τα πιθανά ζεύγη κατηγοριών χρησιμοποιώντας τους προτεινόμενους τύπους δεδομένων εισόδου (λέξεις, λέξεις + λέξεις 2-γραμμα, λέξεις + λέξεις 2 γραμμα + λέξεις 3 γραμμα, λέξεις + λέξεις 2

γραμμα + λέξεις 3 γραμμα + λέξεις 4- γραμμα και λέξεις + λέξεις 2- γραμμα + λέξεις 3- γραμμα + λέξεις 4- γραμμα + λέξεις 5- γραμμα) και τις εφαρμογές του LibSVM και του LR ως ταξινομητές. Να σημειωθεί ότι δεν πραγματοποιήθηκε πείραμα χρησιμοποιώντας την ίδια κατηγορία και για την κατάρτιση και για τις δοκιμές. Για τα πειράματα που αναφέρθηκαν παραπάνω, η προσέγγιση char 3-grams (Stamatatos 2013) χρησιμοποιείται ως γραμμή βάσης και τα αποτελέσματα που προκύπτουν από την εφαρμογή της σε κάθε κατηγορία αντιστοίχισης παρουσιάζονται μαζί με τα αποτελέσματα της προτεινόμενης μεθόδου στους παρακάτω πίνακες.

Ο Πίνακας 2 δείχνει την ακρίβεια που επιτυγχάνεται όταν χρησιμοποιείται η κατηγορία Society(Κοινωνία) για εκπαίδευση στα διάφορα ζεύγη, ο πίνακας 3 δείχνει την ακρίβεια που επιτυγχάνεται όταν χρησιμοποιείται η κατηγορία του Ηνωμένου Βασιλείου για εκπαίδευση στα διάφορα ζεύγη, ο πίνακας 4 δείχνει την ακρίβεια που επιτυγχάνεται όταν χρησιμοποιείται η κατηγορία World(Κόσμος) για εκπαίδευση στα διάφορα ζεύγη και ο Πίνακας 5 δείχνει την ακρίβεια που επιτυγχάνεται όταν χρησιμοποιείται η κατηγορία Books(Βιβλιογραφική κριτική) για εκπαίδευση στα διάφορα ζεύγη. Τα αποτελέσματα που παρουσιάζονται σε αυτούς τους πίνακες υποδεικνύουν ότι η μέθοδος Doc2vec επιτυγχάνει καλύτερη αναπαράσταση κατανομής εγγράφων όταν ο τύπος δεδομένων εισόδου βασίζεται σε λέξεις 1-γράμμων + 2 γράμμων στις περισσότερες περιπτώσεις. Σε μερικές περιπτώσεις, η προσθήκη λέξεων 3-γράμμων, 4-γράμμων ή 5-γράμμων βελτιώνει επίσης τα αποτελέσματα, αλλά με βάση το πείραμά των Posadas Duran et al θωρείται ότι η χρήση λέξεων 1-γράμμων και 2-γράμμων αρκεί. Ειδικά θεωρώντας ότι οι πληροφορίες αυτές διαβιβάζονται στον αλγόριθμο Doc2vec, τόσο περισσότερο χρειάζεται για να μάθει την κατανεμημένη αναπαράσταση.

Name	Politics	UK	World	Books	Average Acc.
D2V words	64.28	55.55	**60.68**	57.14	59.41
D2V words+2-grams	**65.17**	**61.11**	63.24	**53.96**	**53.96**
D2V words+2+3-grams	**65.17**	60.00	63.24	53.96	60.59
D2V words+2+3+4-grams	**64.28**	60.00	63.24	**53.96**	60.37
D2V words+2+3+4+5-grams	**64.28**	60.00	63.24	**53.96**	60.37
Char 3-grams (from Stamatatos 2013)	52.67	45.55	47.00	30.15	43.84
Με έντονη γραφή η υψηλότερη τιμή που επετεύχθη σε κάθε στήλη					

Πίνακας 2. Αποτελέσματα του corpus Guardian υπολογίζοντας ως δεδομένα εκπαίδευσης των φάκελο Society και ως δεδομένα τεστ όλους τους υπόλοιπους φάκελους δεδομένων.

Name	Politics	Society	World	Books	Average Acc.
D2V words	78.57	90.32	78.63	82.53	82.51
D2V words+2-grams	**87.50**	93.54	**83.76**	88.88	88.42
D2V words+2+3-grams	**87.50**	**95.16**	82.90	**90.47**	**89.00**
D2V words+2+3+4-grams	**87.50**	**95.16**	82.90	88.88	88.61
D2V words+2+3+4+5-grams	**87.50**	**95.16**	82.90	88.88	88.61
Char 3-grams (from Stamatatos 2013)	69.64	70.96	60.68	61.90	65.79
Με έντονη γραφή η υψηλότερη τιμή που επετεύχθη σε κάθε στήλη					

Πίνακας 3. Αποτελέσματα του corpus Guardian υπολογίζοντας ως δεδομένα εκπαίδευσης των φάκελο UK και ως δεδομένα τεστ όλους τους υπόλοιπους φάκελους δεδομένων.

Name	Politics	Society	UK	Books	Average Acc.
D2V words	78.57	87.09	80.00	**77.77**	80.85
D2V words+2-grams	83.92	**93.54**	**83.33**	**77.77**	**84.64**
D2V words+2+3-grams	**86.60**	**93.54**	**83.33**	74.60	84.51
D2V words+2+3+4-grams	85.71	**93.54**	82.22	74.60	84.01
D2V words+2+3+4+5-grams	85.71	**93.54**	82.22	74.60	84.01
Char 3-grams (from Stamatatos 2013)	76.78	72.58	63.33	57.14	67.45
Με έντονη γραφή η υψηλότερη τιμή που επετεύχθη σε κάθε στήλη					

Πίνακας 4. Αποτελέσματα του corpus Guardian υπολογίζοντας ως δεδομένα εκπαίδευσης των φάκελο World και ως δεδομένα τεστ όλους τους υπόλοιπους φάκελους δεδομένων.

Name	Politics	Society	UK	World	Average Acc.
D2V words	56.25	51.61	**62.22**	**51.28**	55.34
D2V words+2-grams	**60.71**	**59.67**	**62.22**	48.71	**57.82**
D2V words+2+3-grams	58.92	58.06	60.00	47.00	55.99
D2V words+2+3+4-grams	58.92	58.06	60.00	47.00	55.99
D2V words+2+3+4+5-grams	58.92	58.06	60.00	47.00	55.99
Char 3-grams (from Stamatatos 2013)	43.75	24.19	36.66	32.47	34.26
Με έντονη γραφή η υψηλότερη τιμή που επετεύχθη σε κάθε στήλη					

Πίνακας 5. Αποτελέσματα του corpus Guardian υπολογίζοντας ως δεδομένα εκπαίδευσης των φάκελο Book reviews και ως δεδομένα τεστ όλους τους υπόλοιπους φάκελους δεδομένων.

Name	Politics	Society	UK	World	Books	**Average Acc.**
D2V words	91.68	59.41	82.51	80.85	55.34	73.95
D2V words+2-grams	**94.75**	**60.87**	88.42	**84.64**	**57.82**	**77.30**
D2V words+2+3-grams	93.80	60.59	**89.00**	84.51	55.99	76.77
D2V words+2+3+4-grams	94.20	60.37	88.61	84.01	55.99	65.43
D2V words+2+3+4+5-grams	94.20	60.37	88.61	84.01	55.99	65.43
Char 3-grams (from Stamatatos 2013)	85.50	43.84	65.79	67.45	34.26	59.36
Typed char n-grams (Sapkota et al. 2015)						57.00

Πίνακας 6. Μέσο όρο αποτελεσμάτων του Guardian corpus για κάθε φάκελο δεδομένων.

Author	Politics	Society	World	UK	Reviews
CB	10	4	10	10	10
GM	6	3	10	3	0
HY	8	6	10	5	3
JF	9	1	10	10	2
MK	7	0	10	3	2
MR	8	10	10	10	4
NC	10	2	9	7	5
PP	10	1	10	10	10
PT	10	10	10	5	4
RH	10	4	3	10	10
SH	10	5	5	6	2
WH	10	6	10	5	7
ZW	4	10	10	6	4
Totals	112	62	117	90	63

Σχήμα 1 Το Corpus Guardian για την κλειστή καταχώριση συγγραφέων, με το πολύ 10 έγγραφα ανά συγγραφέα[4]

Έγινε σαφώς εκτενέστερη αναφορά στη μέθοδο που ακολούθησαν οι Posadas Duran et al, καθώς στην επόμενη ενότητα περιγράφουμε τη μέθοδο προσέγγισης μας για την επίλυση της ανάθεσης συγγραφέα χρησιμοποιώντας μια κατανεμημένη αναπαράσταση σε επίπεδο εγγράφου σχετικά όμοια με αυτήν.

[4] **'Application of the distributed document representation in the authorship attribution task for small corpora' Posadas –Duran 2016**

Κεφάλαιο 2

Ανάθεση συγγραφέα χρησιμοποιώντας κατανεμημένη αναπαράσταση εγγράφων

2.1 Κατανεμημένη αναπαράσταση των εγγράφων

Για να βρούμε την κατανεμημένη αναπαράσταση των εγγράφων, δηλ. τα διανύσματα των εγγράφων, χρησιμοποιούμε τη μέθοδο Doc2vec εμπνευσμένη από προηγούμενη έρευνα για την εκμάθηση ενσωμάτωσης λέξεων (Le και Mikolov 2014).

Αρχικά, εισάγουμε την έννοια των κατανεμημένων διανυσμάτων αναπαράστασης των λέξεων. Για το σκοπό αυτό, χρησιμοποιούμε τη μέθοδο που προτείνεται από τους Mikolov et al. (2013α, β). Ο στόχος είναι να προβλέψουμε μια λέξη δεδομένου του πλαισίου που περιβάλλει τη λέξη. Σε αυτή τη μέθοδο, κάθε λέξη χαρτογραφείται σε ένα μοναδικό διάνυσμα που αντιπροσωπεύεται από μια στήλη σε έναν πίνακα W. Επίσημα, δεδομένης μιας ακολουθίας λέξεων εκπαίδευσης w1, w2, w3, ..., wT, ο στόχος της εκπαίδευσης είναι να μεγιστοποιηθεί ο μέσος όρος της πιθανότητας log -:

Η εργασία πρόβλεψης εκτελείται από έναν πολυταξικό ταξινομητή, όπως ο Softmax:

όπου yi είναι το i-στοιχείο του διανύσματος του y-class. Ο παραπάνω τύπος μπορεί να ερμηνευτεί ως η (κανονικοποιημένη) πιθανότητα που αποδίδεται στη σωστή ετικέτα wt με δεδομένες τις λέξεις εκπαίδευσης wt-k, ..., wt + k. Κάθε yi είναι μη κανονικοποιημένη λογαριθμική πιθανότητα για κάθε λέξη έξοδου i και υπολογίζεται ως:

$$y = b + Uh(wt-k , ...,wt+k ;W), \qquad (1)$$

όπου U και b είναι οι παράμετροι Softmax. το h κατασκευάζεται με τη συνένωση ή τον μέσο όρο των διανυσμάτων των λέξεων που εξάγονται από το W.

Για να πάρουμε την λογαριθμική πιθανότητα καταγραφής για κάθε λέξη χρησιμοποιείται ένας υπολογιστικός αποδοτικός αλγόριθμος για την διαδικασία πρόβλεψης: ο ιεραρχικός Softmax (Mnih και Hinton 2008, Mikolov et al., 2013b). Η δομή της ιεραρχίας του Softmax είναι ένα δυαδικό δέντρο Huffman (Bird and Wadler

1988), όπου μικροί κωδικοί ανατίθενται στις συχνές λέξεις. Μετά τη σύγκλιση της εκπαίδευσης, οι λέξεις με παρόμοια σημασία χαρτογραφούνται σε μια παρόμοια θέση στο διανυσματικό χώρο (Mikolov et al., 2013c).

Για να μάθουμε τα διανύσματα των παραγράφων, ακολουθείται η ίδια μέθοδος. Τα διανύσματα παραγράφων καλούνται να συμβάλλουν στην εργασία πρόβλεψης της επόμενης λέξης που δίνεται σε πολλά περιστατικά που έχουν δειγματοληψία από το έγγραφο με τον ίδιο τρόπο με τον οποίο τα διανύσματα λέξεων καλούνται να συνεισφέρουν σε μια πρόβλεψη για την επόμενη λέξη σε μια πρόταση. Τα διανύσματα των λέξεων ή των παραγράφων αρχικοποιούνται τυχαία, αλλά στο τέλος, καταγράφουν τη σημασιολογία ως έμμεσο αποτέλεσμα της πρόβλεψης. Τέλος, τα διανύσματα για τα έγγραφα λαμβάνονται με παρόμοιο τρόπο (Mikolov et al., 2013c).

Στο μοντέλο διανυσμάτων εγγράφων, κάθε έγγραφο χαρτογραφείται σε ένα μοναδικό διάνυσμα που αντιπροσωπεύεται από μια στήλη στον πίνακα D με τον ίδιο τρόπο: Κάθε λέξη χαρτογραφείται σε ένα μοναδικό διάνυσμα που αντιπροσωπεύεται από μια στήλη στον πίνακα W. Στο μοντέλο διανυσμάτων εγγράφων, Εξ. 1 μεταβάλλεται για να κατασκευαστεί το h από το W και το D.

Υπάρχουν δύο μοντέλα για κατανεμημένη αναπαράσταση εγγράφων: κατανεμημένη μνήμη (DM) και κατανεμημένη σακούλα λέξεων (DBOW). Στην περίπτωση της DM, κάθε παράγραφος είναι ένα διάνυσμα των τιμών των χαρακτηριστικών, τα οποία χρησιμοποιούνται για την πρόβλεψη των διανυσμάτων των παραγράφων πλαισίου (προηγούμενες ή επόμενες παραγράφους). Το DBOW αγνοεί τις λέξεις πλαισίου (ή n-grams στην περίπτωσή μας) στην είσοδο, αλλά προβλέπει τις λέξεις που δειγματοληπτούντε τυχαία από τις παραγράφους στην έξοδο. Έτσι, σε κάθε επανάληψη της κατάταξης στοχαστικής κλίσης ερευνά ένα παράθυρο κειμένου και εκτελεί μια εργασία ταξινόμησης δεδομένου του διανύσματος παραγράφου.

2.2 Απόδοση συγγραφέα ως πρόβλημα ταξινόμησης

Υπάρχουν δύο τύποι: κλειστός, όταν το σύνολο συγγραφέων είναι προκαθορισμένο και ανοιχτό, όταν ένας νέος συγγραφέας (που απουσιάζει από το σώμα εκπαίδευσης) μπορεί να παρουσιαστεί στη μέθοδο. Εμείς, ασχολούμαστε με την κλειστή κατανομή συγγραφέων. Η εφαρμογή της κατανομής κλειστού συγγραφικού περιεχομένου μπορεί να θεωρηθεί ως ένα πρόβλημα πολλαπλών κλάσεων ταξινόμησης που ορίζεται ως εξής:

Λαμβάνοντας υπόψη ένα σύνολο γνωστών συγγραφέων A = {A1, A2, ..., Ai} και ένα σύνολο παραδειγμάτων T = t11, t12,. . . , t1n,. . . , tij, όπου το στοιχείο tij αντιστοιχεί στο j παράδειγμα του συγγραφέα Ai, το πρόβλημα είναι να δημιουργήσουμε έναν ταξινομητή F ο οποίος αναθέτει κάθε στοιχείο ενός συνόλου κειμένων άγνωστης συγγραφικής ταυτότητας X = {x1, x2,. . . , xm} σε έναν μόνο γνωστό συγγραφέα, δηλ. F: X → A.

Η προσέγγισή μας για την ανάθεση συγγραφέων βασίζεται στη μηχανική μάθηση, η οποία αποτελείται από δύο φάσεις: τη φάση της κατάρτισης και τη φάση της δοκιμής. Στη φάση της κατάρτισης μαθαίνουμε την κατανεμημένη αναπαράσταση των εγγράφων στο σύνολο T, δηλαδή V = v11, v12,. . . , tij, όπου vij = {f1, f2,. . . , fn} είναι η διανυσματική παράσταση του κειμένου ti1j. Η αντιπροσώπευση του διανύσματος επιτυγχάνεται με την εφαρμογή της μεθόδου Doc2vec (Le και Mikolov 2014) στο GENSIM.1

Η μέθοδος Doc2vec δημιουργεί ένα μοντέλο για τη λήψη κατανεμημένων αναπαραστάσεων εγγράφων. προσφέρει δύο πιθανές προσεγγίσεις για την κατασκευή του μοντέλου: το DM και το DBOW όπως αναφέρθηκε προηγουμένως. Προηγούμενη έρευνα για την ανάλυση του συναισθήματος αναφέρει καλύτερα αποτελέσματα όταν συνδυάζονται και οι δύο παραστάσεις, οπότε στην πρότασή μας το τελικό διάνυσμα εγγράφων αποτελείται από τη συνένωση των παραστάσεων που αποκτήθηκαν από τα μοντέλα DM και DBOW. Θα δούμε αναλυτικά το doc2vec παρακάτω.

Ένας ταξινομητής εκπαιδεύεται με τις αναπαραστάσεις των διανυσμάτων των εγγράφων χρησιμοποιώντας μια πολύ γνωστή μέθοδο, την SVM. Επιλέξαμε αυτόν τον ταξινομητή επειδή είναι ισχυρός με αραιά δεδομένα και μπορεί να χειριστεί βέλτιστα διανύσματα με μεγάλη διαστασιολόγηση.

Στη φάση της δοκιμής, παίρνουμε τις αναπαραστάσεις των διανυσμάτων των κειμένων του άγνωστου συντάκτη. Διαμορφώνεται με τη συνένωση των παραστάσεων που αποκτώνται με τα μοντέλα Doc2vec που χρησιμοποιούνται στη φάση της εκπαίδευσης. Τέλος, ο ταξινομητής αναθέτει τον συγγραφέα (κλάση) σε κάθε έγγραφο χρησιμοποιώντας αυτές τις κατανεμημένες αναπαραστάσεις (διανύσματα) ως χαρακτηριστικά.

Κεφάλαιο 3

Το Doc2vec

3.1 Εισαγωγή στο Doc2Vec

Ας δούμε όμως τι είναι το doc2vec , πώς είναι υλοποιημένο , πώς σχετίζεται με word2vec και τι μπορούμε να κάνουμε με αυτό,

Η αριθμητική αναπαράσταση εγγράφων κειμένου είναι ένα δύσκολο έργο στη μηχανική μάθηση. Μια τέτοια αναπαράσταση μπορεί να χρησιμοποιηθεί για πολλούς σκοπούς, όπως: ανάκτηση εγγράφων, αναζήτηση ιστού, φιλτράρισμα ανεπιθύμητων μηνυμάτων, μοντελοποίηση θέματος κ.λπ.

Ωστόσο, δεν υπάρχουν πολλές καλές τεχνικές για να γίνει αυτό. Πολλές εργασίες χρησιμοποιούν τη γνωστή αλλά απλοϊκή μέθοδο (BOW) , αλλά τα αποτελέσματα είναι συνήθως μέτρια, αφού το BOW χάνει πολλές λεπτομέρειες.

Το Latent Dirichlet Allocation (LDA) είναι επίσης μια κοινή τεχνική (εξάγοντας θέματα / λέξεις-κλειδιά από κείμενα) αλλά είναι πολύ δύσκολο τα αποτελέσματα να αξιολογηθούν.

Το doc2vec , παρουσιάστηκε όπως αναφέραμε το 2014 από τους Mikolov και Le Αξίζει να αναφέρουμε ότι ο Mikolov είναι ένας από τους συγγραφείς του word2vec .

Το Doc2vec είναι μια πολύ ωραία τεχνική. Είναι εύκολο στη χρήση, δίνει καλά αποτελέσματα, και όπως μπορείτε να καταλάβετε από το όνομά του, βασίζεται στο word2vec. Οπότε θα κάνουμε και μια σύντομη εισαγωγή για word2vec.

3.2 word2vec

Το word2vec είναι αρκετά γνωστό , και χρησιμοποιείται για τη δημιουργία διανυσμάτων αναπαράστασης από λέξεις.

Υπάρχουν πολλά καλά σεμινάρια online σχετικά με word2vec, , αλλά περιγράφοντας το doc2vec χωρίς να αναφέρουμε στο word2vec θα χαθούν σημαντικές λεπτομέρειες

Σε γενικές γραμμές, όταν θέλουμε να δημιουργήσουμε κάποιο μοντέλο χρησιμοποιώντας λέξεις, απλά η επισήμανση / η κωδικοποίηση αποτελεί έναν εύκολο τρόπο. Ωστόσο, με μια τέτοια κωδικοποίηση, οι λέξεις χάνουν το νόημά τους.

Η word2vec, που παρουσιάστηκε το 2013, σκοπεύει να δώσει ακριβώς αυτό: μια αριθμητική αναπαράσταση για κάθε λέξη, που θα μπορέσει να συλλάβει τις σχέσεις μεταξύ των λέξεων. Αυτό είναι μέρος μιας ευρύτερης έννοιας στη μηχανική μάθηση και στα διανύσματα χαρακτηριστικών.

Τέτοιες αναπαραστάσεις, ενσωματώνουν διαφορετικές σχέσεις μεταξύ λέξεων, όπως συνώνυμα, αντώνυμα ή αναλογίες.

3.3 Αλγόριθμοι Word2vec

Το word2vec δημιουργείται χρησιμοποιώντας 2 αλγόριθμους: Το Μοντέλο (CBOW) και το μοντέλο Skip-Gram .

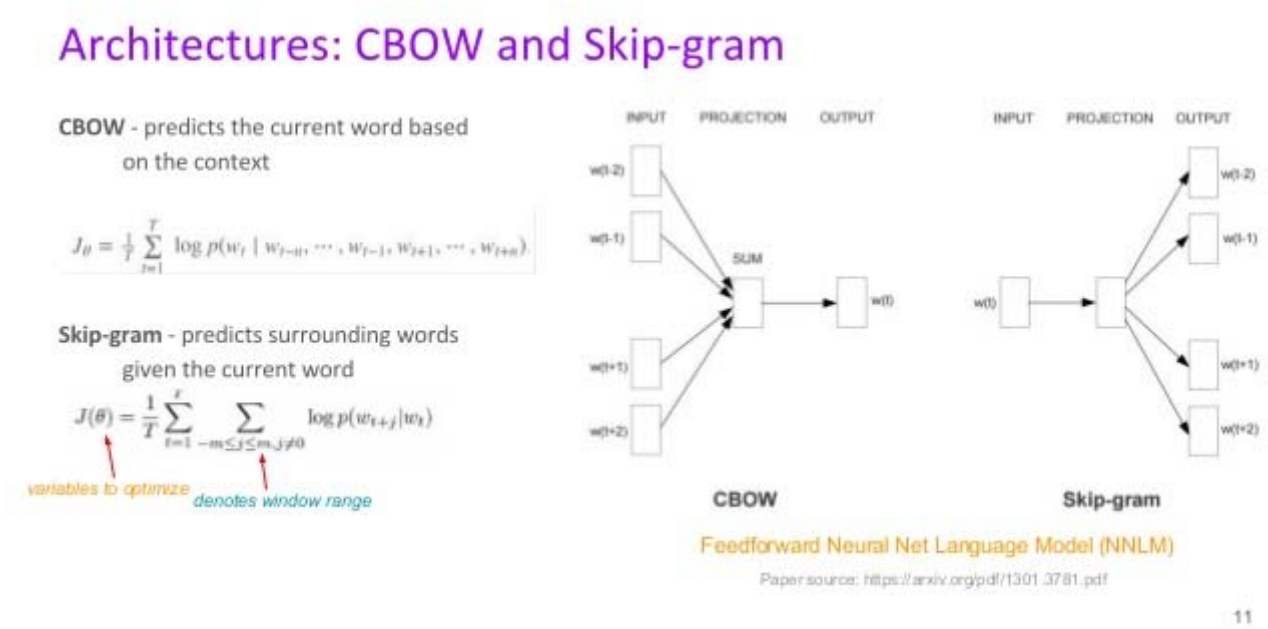
3.3.1 CBOW

Το cbow δημιουργεί ένα παράθυρο γύρω από την τρέχουσα λέξη, για να την προβλέψει από τα γύρω λόγια. Κάθε λέξη αντιπροσωπεύεται ως διάνυσμα χαρακτηριστικών. Μετά την εκπαίδευση, αυτά τα διανύσματα γίνονται οι φορείς λέξεων.

Όπως αναφέρθηκε προηγουμένως, οι φορείς που αναπαριστούν παρόμοιες λέξεις είναι κοντά σε μετρικές αποστάσεις.

3.3.2 Το Skip gram

Ο δεύτερος αλγόριθμος είναι αντίθετα από το CBOW: αντί να προβλέπουμε μία λέξη κάθε φορά, χρησιμοποιούμε 1 λέξη για να προβλέψουμε όλες τις λέξεις γύρω από το περιβάλλον. Το Skip gram είναι πολύ πιο αργό από το CBOW, αλλά θεωρείται πιο ακριβές με σπάνιες λέξεις.



Εικόνα 2 Αρχιτεκτονική του doc2vec[5]

[5] https://www.slideshare.net/ibelmopan/text-mining-lab-summer-2017-word-vector-representation

3.4 Doc2vec

Αφού είδαμε λίγο το word2vec, θα είναι ευκολότερο να καταλάβουμε πώς λειτουργεί το doc2vec .

Όπως είπαμε, ο στόχος του doc2vec είναι να δημιουργήσει μια αριθμητική αναπαράσταση ενός εγγράφου, ανεξάρτητα από το μήκος του. Αλλά σε αντίθεση με τα λόγια, τα έγγραφα δεν έρχονται σε λογικές δομές, οπότε πρέπει να βρεθεί και μια άλλη μέθοδος.

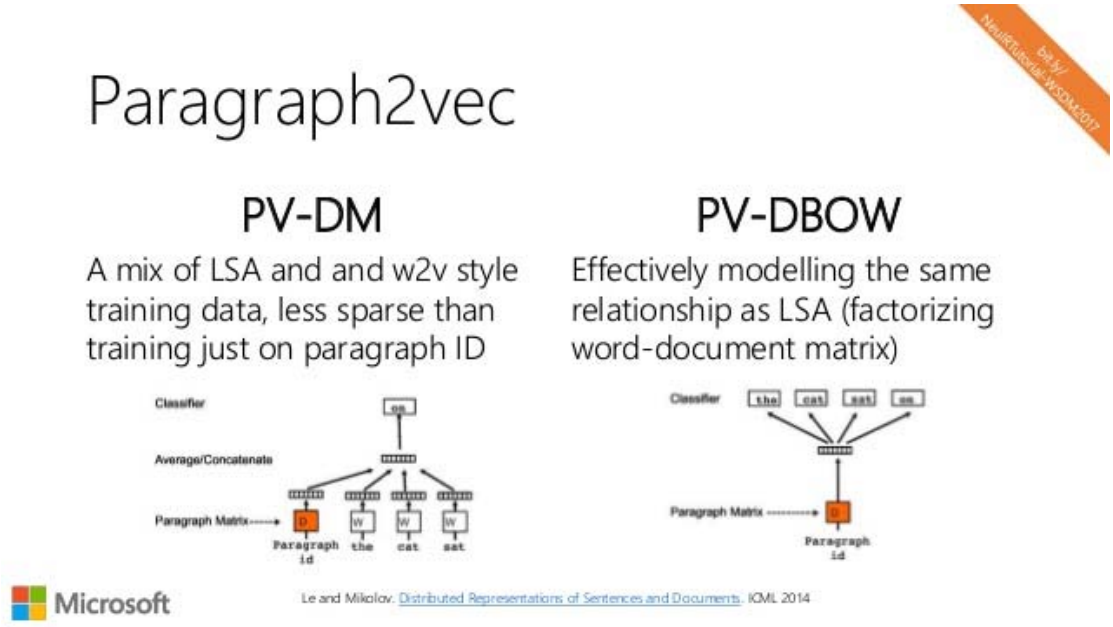
Η ιδέα που χρησιμοποίησαν οι Mikolov και Le ήταν απλή. Χρησιμοποίησαν το μοντέλο word2vec και πρόσθεσαν ένα ακόμη διάνυσμα .

Αντί να χρησιμοποιήσουμε μόνο λέξεις για να προβλέψουμε την επόμενη λέξη, προστέθηκε επίσης ένα άλλο διάνυσμα χαρακτηριστικών, το οποίο είναι μοναδικό για τα έγγραφα.

Έτσι, κατά την εκπαίδευση των λέξεων vectors W, το διάνυσμα εγγράφων D εκπαιδεύεται επίσης, και στο τέλος της εκπαίδευσης, κατέχει μια αριθμητική αναπαράσταση του εγγράφου.

Το παραπάνω μοντέλο ονομάζεται έκδοση κατανεμημένης μνήμης Vector (PV-DM). Λειτουργεί ως μνήμη που θυμάται τι λείπει από το τρέχον πλαίσιο Ενώ οι φορείς λέξης αντιπροσωπεύουν την έννοια μιας λέξης, το διάνυσμα εγγράφων σκοπεύει να αντιπροσωπεύει την έννοια ενός εγγράφου.

Όπως και στο word2vec, ένας άλλος αλγόριθμος, ο οποίος είναι παρόμοιος με το skip-gram, μπορεί να χρησιμοποιηθεί, ο Distributed Bag of Words έκδοση του Vector Paragraph (PV-DBOW)



Εικόνα 3 Paragraph2vec[6]

Εδώ, ο αλγόριθμος αυτός είναι στην πραγματικότητα γρηγορότερος και καταναλώνει λιγότερη μνήμη, αφού δεν χρειάζεται να αποθηκεύει διανύσματα λέξεων.

Οι συγγραφείς δηλώνουν ότι συνιστούν τη χρήση ενός συνδυασμού και των δύο αλγορίθμων, αν και το μοντέλο PV-DM είναι ανώτερο και συνήθως επιτυγχάνει τα αποτελέσματα από μόνο του.

Τα μοντέλα doc2vec μπορούν να χρησιμοποιηθούν με τον ακόλουθο τρόπο:

- Για την κατάρτιση απαιτείται ένα σύνολο εγγράφων.
- Για κάθε λέξη δημιουργείται ένας φορέας λέξεων W και για κάθε έγγραφο δημιουργείται ένας φορέας εγγράφων D.
- Το μοντέλο επίσης εκπαιδεύει τα βάρη για ένα κρυφό στρώμα softmax.
- Στο στάδιο συμπερασμάτων, μπορεί να παρουσιαστεί ένα νέο έγγραφο και όλα τα βάρη έχουν καθοριστεί για τον υπολογισμό του διανύσματος εγγράφων.

[6] https://www.slideshare.net/BhaskarMitra3/neural-text-embeddings-for-information-retrieval-wsdm-2017

3.5 Αξιολόγηση

Αυτό το είδος μη εποπτευόμενων μοντέλων είναι ότι δεν εκπαιδεύονται για να κάνουν το έργο που προορίζονται. Π.χ., το word2vec εκπαιδεύεται για να ολοκληρώσει τις λέξεις γύρω από το σώμα, και χρησιμοποιείται για να υπολογίσει την ομοιότητα ή τις σχέσεις μεταξύ των λέξεων. Ως εκ τούτου, η μέτρηση της απόδοσης αυτών των αλγορίθμων μπορεί να είναι δύσκολη.

Επομένως, κατά την κατάρτιση αυτών των αλγορίθμων, θα πρέπει να είμαστε προσεκτικοί στις σχετικές μετρήσεις[14][15].

Κεφάλαιο 4

Χρήση του Doc2Vec

Το μοντέλο Doc2vec από μόνο του είναι μια μη επιτηρούμενη μέθοδος, οπότε πρέπει να τροποποιηθεί λίγο για να μπορέσουμε να το χρησιμοποιήσουμε . Ευτυχώς, όπως στις περισσότερες περιπτώσεις, μπορούμε να κάνουμε κάποιες αλλαγές. Μπορούμε να προσθέσουμε ένα άλλο διάνυσμα εγγράφων, μοναδικό για κάθε έγγραφο. Έτσι μπορούμε να προσθέσουμε πολλούς συγγραφείς οι οποίοι δεν πρέπει να είναι μοναδικοί: για παράδειγμα, εάν έχουμε ετικέτες για τα έγγραφά μας (όπως έχουμε στην πραγματικότητα), μπορούμε να τα προσθέσουμε και να πάρουμε την εκπροσώπησή τους ως φορείς.

Επιπλέον, δεν χρειάζεται να είναι μοναδικά. Με αυτόν τον τρόπο, μπορούμε να προσθέσουμε στην ετικέτα του μοναδικού εγγράφου μία από τις 17 ετικέτες μας και να δημιουργήσουμε μια αντιπροσωπεία doc2vec και για αυτούς!

Θα χρησιμοποιήσουμε το gensim στο doc2vec. Το gensim Tagged Document και στη συνέχεια μπορούμε να ελέγξουμε την ομοιότητα κάθε μοναδικού εγγράφου σε κάθε ετικέτα με αυτόν τον τρόπο:

Η ετικέτα με την υψηλότερη ομοιότητα με το έγγραφο θα προβλεφθεί και θα παρουσιαστεί..

4.1 Περίληψη

Έχουμε δει ότι με κάποιες μικρό-αλλαγές, μπορούμε να πάρουμε πολύ περισσότερα από το ήδη πολύ χρήσιμο μοντέλο word2vec . Αυτό είναι σπουδαίο, διότι, όπως είπαμε προηγουμένως, κατά τη γνώμη μου η παρουσίαση των εγγράφων για την επισήμανση και την αντιστοίχιση σε συγγραφέα μπορεί να βελτιωθεί.

Στόχος του προγράμματος είναι η σύγκριση 2 κειμένων του ενός εκ των οποίων γνωρίζουμε τον συγγραφέα. Προσπαθήσαμε να βρούμε το ποσοστό στο οποίο το άγνωστο κείμενο ενδέχεται να είναι του συγγραφέα. Το πρόγραμμα υλοποιείται σε γλώσσα python . Οι τεχνολογίες που θα ενσωματωθούν είναι τα n-grams και το doc2vec.

Το πρόγραμμα παίρνει ένα κείμενο και το σπάει αρχικά σε προτάσεις. Έτσι δημιουργείται ένα μοντέλο με βάση το doc2vec και στη συνέχεια οι προτάσεις σπάνε με βάση τα n-grams ανάλογα με τη μεταβλητή (n) που υπάρχει στο πρόγραμμα. Καταφέραμε να φτάσουμε το ποσοστό επιτυχίας του αλγορίθμου μέχρι και 60 % περίπου καθώς όπως είπαμε ο αλγόριθμος μαθαίνει από τα κείμενα και σε μας ήταν μικρό το δείγμα που χρησιμοποιήσαμε και για αυτό το λόγο θεωρείται ικανοποιητικό.

Τα n-grams των 2 κειμένων ελέγχονται για ταυτίσεις και το ποσοστό βγαίνει βάση αυτής της σύγκρισης.

Επιπλέον, αυτό δείχνει ότι αυτό είναι ένα καλό παράδειγμα για το πώς τα μοντέλα μηχανικής μάθησης ενσωματώνουν πολύ περισσότερες δυνατότητες εκτός από το συγκεκριμένο έργο για το οποίο εκπαιδεύτηκαν. Αυτό μπορεί να φανεί σε βαθιά CNN, τα οποία είναι εκπαιδευμένα για την ταξινόμηση αντικειμένων, αλλά μπορούν επίσης να χρησιμοποιηθούν για σημασιολογική κατάτμηση ή ομαδοποίηση εικόνων.

4.2 Ανάλυση

Η τελευταία έκδοση gensim έχει μια νέα κλάση που ονομάζεται Doc2Vec . Όλες οι αξιολογήσεις για αυτή την τάξη, η οποία είναι μια εφαρμογή του Quoc Le & Tomáš Mikolov: "Κατανεμημένες Αντιπροσωπείες Σημείων και Εγγράφων" , καθώς και αυτό το κείμενο, βασίζεται στην ανάλυση του Tim Emerick.

Το Doc2vec (γνωστός και ως paragraph2vec, γνωστός ως "embeddings sentence") τροποποιεί τον αλγόριθμο word2vec στην εκμάθηση συνεχών

αναπαραστάσεων για μεγαλύτερα τμήματα κειμένου , όπως προτάσεις, παραγράφους ή ολόκληρα έγγραφα.

Η λειτουργικότητα του doc2vec αναβαθμίστηκε στο gensim 0.12.0. Το API είναι πλέον καθαρότερο, εκπαιδεύεται γρηγορότερα, υπάρχουν περισσότερες παράμετροι ρύθμισης που έχουν εκτεθεί κλπ. Ενώ οι βασικές ιδέες που εξηγούνται παρακάτω ισχύουν,. Για μια μηχανή ομοιότητας εμπορικών εγγράφων, ανατρέξτε στο scaletext.com .

4.2.1 Εισαγωγή

Δεδομένου ότι η κλάση Doc2Vec επεκτείνει την αρχική κατηγορία Word2Vec της gensim , πολλά από τα πρότυπα χρήσης είναι παρόμοια. Μπορούμε να ρυθμίσουμε εύκολα τη διάσταση της παράστασης, το μέγεθος του παραθύρου, τον αριθμό των εγγράφων ή σχεδόν οποιαδήποτε άλλη παράμετρο που μπορούμε να αλλάξουμε με το μοντέλο Word2Vec.

Η μόνη εξαίρεση είναι οι παράμετροι που σχετίζονται με τη μέθοδο εκπαίδευσης που χρησιμοποιείται από το μοντέλο. Στην αρχιτεκτονική word2vec, τα δύο ονόματα αλγορίθμων είναι (cbow) και "skip-gram" (sg). στην αρχιτεκτονική doc2vec, οι αντίστοιχοι αλγόριθμοι είναι "κατανεμημένη μνήμη" (dm) και "distributed bag of words" (dbow). Δεδομένου ότι το μοντέλο κατανεμημένης μνήμης εκτελέστηκε αισθητά καλύτερα στο χαρτί, αυτός ο αλγόριθμος είναι ο προεπιλεγμένος κατά την εκτέλεση του Doc2Vec. Μπορούμε ακόμα να χρησιμοποιήσουμε το μοντέλο dbow εάν θέλουμε, χρησιμοποιώντας dm=0 στον κατασκευαστή.

Η είσοδος στο Doc2Vec είναι ένα iterator αντικειμένων LabeledSentence . Κάθε τέτοιο αντικείμενο αντιπροσωπεύει μια μόνο πρόταση και αποτελείται από δύο απλές λίστες: μια λίστα λέξεων και μια λίστα ετικετών :

Ο αλγόριθμος τρέχει τότε μέσω των iterator προτάσεων δύο φορές: μία φορά για να δημιουργήσει το vocab, και μία φορά για να εκπαιδεύσει το μοντέλο στα δεδομένα εισόδου, μαθαίνοντας μια αναπαράσταση του φορέα για κάθε λέξη και για κάθε ετικέτα στο σύνολο δεδομένων.

Αν και αυτή η αρχιτεκτονική επιτρέπει περισσότερες από μία ετικέτες ανά πρόταση η πιο δημοφιλής περίπτωση χρήσης θα ήταν να υπάρχει μια ενιαία ετικέτα ανά πρόταση, η οποία είναι το μοναδικό αναγνωριστικό για την πρόταση.

Μια πιο ισχυρή έκδοση αυτής της κατηγορίας LabeledLineSentence παραπάνω περιλαμβάνεται επίσης στο module doc2vec , έτσι μπορείτε να το χρησιμοποιήσετε. Διαβάστε τα έγγραφα doc2vec API για όλες τις παραμέτρους του κατασκευαστή.

4.2.2 Εκπαίδευση

Το Doc2Vec μαθαίνει παραστάσεις για λέξεις και ετικέτες ταυτόχρονα. Αν θέλουμε να μάθει μόνο παραστάσεις για λέξεις, μπορούμε να χρησιμοποιήσουμε τη σημαία train_lbls=False στην κλάση Doc2Vec. Ομοίως, εάν επιθυμούμε μόνο να μάθει παραστάσεις για ετικέτες και να αφήσει τις εκφράσεις λέξεων σταθερές, το μοντέλο έχει επίσης τη σημαία train_words=False.

Μία παρατήρηση του τρόπου με τον οποίο ο αλγόριθμος αυτός εκτελείται είναι ότι, καθώς ο ρυθμός εκμάθησης μειώνεται κατά τη διάρκεια της αλλαγής των δεδομένων, οι ετικέτες που φαίνονται μόνο σε ένα μόνο LabeledSentence κατά τη διάρκεια της εκπαίδευσης θα εκπαιδευτούν μόνο με έναν σταθερό ρυθμό εκμάθησης. Αυτό συχνά παράγει λιγότερο από τα βέλτιστα αποτελέσματα.

4.2.3 Χρήση μνήμης

Με την τρέχουσα υλοποίηση, όλοι οι ετικέτες ετικετών αποθηκεύονται ξεχωριστά στη μνήμη RAM. Στην παραπάνω περίπτωση με μια μοναδική ετικέτα ανά πρόταση, αυτό προκαλεί την αύξηση της χρήσης της μνήμης με το μέγεθος του κειμένου, το οποίο μπορεί να είναι ή όχι πρόβλημα, ανάλογα με το μέγεθος του κειμένου σας και το διαθέσιμο RAM στο κουτί σας.

4.2.4 I / O

Η χρήση του Doc2Vec είναι ίδια με αυτή του Word2Vec της gensim. Κάποιος μπορεί να σώσει και να φορτώσει gensim περιπτώσεις Doc2Vec στους συνήθεις τρόπους: άμεσα με Python, ή να χρησιμοποιούν τις βελτιστοποιημένες Doc2Vec.save()και Doc2Vec.load()μεθόδους:

```
model = Doc2Vec(sentences)

...

# store the model to mmap-able files

model.save('/tmp/my_model.doc2vec')

# load the model back

model_loaded = Doc2Vec.load('/tmp/my_model.doc2vec')
```

Οι λειτουργίες του βοηθού σαν model.most_similar(), model.doesnt_match()και model.similarity()επίσης υπάρχουν. Οι ακατέργαστες λέξεις και οι ετικέτες της ετικέτας είναι επίσης προσβάσιμες είτε μεμονωμένα μέσω model['word'], είτε όλοι ταυτόχρονα μέσω model.syn0.

Το κύριο σημείο είναι ότι οι ετικέτες λειτουργούν με τον ίδιο τρόπο όπως οι λέξεις στο Doc2Vec . Έτσι, για να πάρετε τις πιο παρόμοιες λέξεις / προτάσεις στην πρώτη πρόταση (SENT_0για παράδειγμα, την ετικέτα), θα πληκτρολογούσαμε:

```
print model.most_similar(&amp;quot;SENT_0&amp;quot;)

[('SENT_48859', 0.2516525387763977),

 (u'paradox', 0.24025458097457886),

 (u'methodically', 0.2379375547170639),
```

(u'tongued', 0.22196565568447113),

(u'cosmetics', 0.21332012116909027),

(u'Loos', 0.2114654779434204),

(u'backstory', 0.2113303393125534),

('SENT_60862', 0.21070502698421478),

(u'gobble', 0.20925869047641754),

('SENT_73365', 0.20847654342651367)]

Το έγγραφο Doc2vec εισήχθη από τον Le & Mikolov στο έγγραφο ICML'14 - "Distributed Representations of Sentences and Documents". Όπως αναφέρετε, πρόκειται για ένα σύνολο προσεγγίσεων που αντιπροσωπεύουν έγγραφα ως διανύσματα χαμηλής διαστάσεως σταθερού μήκους (επίσης γνωστά ως ενσωματώσεις εγγράφων). Σημειώστε ότι το doc2vec δεν είναι η μόνη προσέγγιση για τη δημιουργία παραστάσεων χαμηλής διάστασης. Ωστόσο, έχουν υπάρξει πρόσφατοι ισχυρισμοί ότι το doc2vec ξεπερνά τα υπόλοιπα συστήματα ενσωμάτωσης.

Για να κατανοήσουμε το doc2vec, καλό είναι να δούμε τη προσέγγιση word2vec, καθώς ο προηγούμενος είναι μια επέκταση του τελευταίου. Οι μέθοδοι με βάση το Word2vec αποσκοπούν στον υπολογισμό των αναπαραστάσεων του φορέα των λέξεων επίσης γνωστές ως ενσωματωμένες λέξεις (embedded words).

Το Word2vec είναι ένα νευρωνικό δίκτυο τριών επιπέδων με μια είσοδο, ένα κρυφό και ένα στρώμα εξόδου. Η ιδέα της αρχιτεκτονικής CBOW ,ένας από τους αλγόριθμους που βασίζονται στο word2vec, είναι να μαθαίνει εκφράσεις λέξεων που μπορούν να προβλέψουν μια λέξη δεδομένης της λέξεις που την περιβάλλουν.

Το στρώμα εισόδου αντιστοιχεί σε σήματα για το περιβάλλον (λέξεις γύρω από το κείμενο) και το επίπεδο εξόδου αντιστοιχεί σε σήματα για την προβλεπόμενη λέξη στόχου .

Πριν προχωρήσουμε στο doc2vec, θα δούμε λίγο την έννοια του "πλαισίου". Η διαδικασία κατάρτισης word2vec, μπορεί να θεωρηθεί ως εποπτευόμενο έργο

εκμάθησης της πρόβλεψης της τάξης λέξεων-στόχων, δεδομένου του πλαισίου εισόδου.. Είναι σημαντικό να κατανοήσουμε ότι το πλαίσιο είναι πιο γενικό από απλά λόγια. Θα μπορούσε να είναι οποιοδήποτε σήμα χρήσιμο στην πρόβλεψη της λέξης στόχου.

Για παράδειγμα, μέρος των ετικετών ομιλίας προηγούμενων λέξεων μπορεί να αποτελεί πλαίσιο. Εάν έχουμε πολλά έγγραφα από τα οποία έρχονται ζεύγη (πλαίσιο, στόχος), το ίδιο το έγγραφο μπορεί να χρησιμεύσει ως πλαίσιο.

Σκεφτείτε το αν γνωρίζετε ότι ένα πρότυπο παράδειγμα "ναι εμείς <στόχος>!" Προέρχεται από μια σελίδα wikipedia σχετικά με τις "αμερικανικές προεδρικές εκλογές του 2008", είναι πολύ πιο εύκολο να προβλέψετε αυτόν τον στόχο ως τη λέξη "μπορεί".

Το Doc2Vec εξετάζει την παραπάνω παρατήρηση προσθέτοντας επιπλέον κόμβους εισόδου που αντιπροσωπεύουν τα έγγραφα ως πρόσθετο πλαίσιο. Κάθε πρόσθετος κόμβος μπορεί να θεωρηθεί ακριβώς ως αναγνωριστικό για κάθε έγγραφο εισόδου.

Τα χαρακτηριστικά D που αντιπροσωπεύουν το περιβάλλον του εγγράφου και το W αντιπροσωπεύουν το πλαίσιο λέξεων σε ένα παράθυρο που περιβάλλει τη λέξη στόχου. Η εκπαίδευση είναι παρόμοια με word2vec, με πρόσθετο περιβάλλον εγγράφων.

Στο τέλος της εκπαιδευτικής διαδικασίας, θα έχετε ενσωματωμένες λέξεις, W και την ενσωμάτωση εγγράφων D για έγγραφα στο σώμα εκπαίδευσης.

Έχουμε ενσωματώσει έγγραφα για το corpus εισόδου αλλά τι γίνεται με τα νέα αόρατα έγγραφα που εμφανίζονται στο σετ δοκιμών; Η ιδέα είναι να μάθουμε την εκπροσώπησή τους στο χρόνο δοκιμής λύνοντας ένα πρόβλημα βελτιστοποίησης για συμπέρασμα.

Το πρόβλημα βελτιστοποίησης δεν διαφέρει από το πρόβλημα της κατάρτισης. Κάποιος μπορεί ωστόσο να επιλέξει να κρατήσει το W και W' σταθερή και να μάθει μεταβλητή D ως ενσωμάτωση εγγράφων.

Προχωρώντας στο δεύτερο μέρος της ερώτησης σχετικά με τα εμπλεκόμενα μαθηματικά, θα επισημανθούν δύο πόροι. Το ένα είναι ιδιαίτερα στο word2vec - "

Εξήγηση μαθηματικών παραμέτρων Word2Vec εξήγησε " ο Xin Rong. Αυτό έχει παραδοχές των εξισώσεων επικαιρότητας word2vec με όλες τις "βαριές" λεπτομέρειες της αρνητικής δειγματοληψίας ή της ιεραρχικής softmax

Οι ενσωματώσεις / διανύσματα λέξεων είναι μια ισχυρή μέθοδος που βοήθησε σε μεγάλο βαθμό τις μεθόδους NLP με βάση τα νευρωνικά δίκτυα. Μέχρι τώρα, το word2vec και το GloVe τείνουν να χρησιμοποιούνται ως η τυπική μέθοδος για την απόκτηση λέξεων ενσωματωμένων λέξεων(embedded words).

Παρόλο που οι ενσωματώσεις φράσεων μπορούν συχνά να αποκτηθούν με εποπτευόμενο τρόπο, με την κατάρτιση ενός μοντέλου πάνω από τις ενσωματωμένες λέξεις, υπάρχουν συχνά περιπτώσεις στις οποίες θέλουμε να αποκτήσουμε ενσωματώσεις φράσεις με ανεξέλεγκτο τρόπο . Για παράδειγμα, μπορεί να θέλουμε να διενεργήσουμε ταυτοποίηση με παράφραση ή να δημιουργήσουμε ένα σύστημα για την αποτελεσματική ανάκτηση παρόμοιων προτάσεων, όπου δεν έχουμε ρητά δεδομένα εποπτείας.

4.3 Θεματικά μοντέλα

Αν και συχνά επισκιάζεται από τις μεθόδους που βασίζονται στα νευρωνικά δίκτυα, η μοντελοποίηση θέματος είναι ένα εκπληκτικά ισχυρό και ερμηνεύσιμο πλαίσιο για την ενσωμάτωση προτάσεων.

Η μοντελοποίηση του θέματος βασικά αποσκοπεί στην αποσύνθεση ενός συνόλου εγγράφων / προτάσεων σε διάφορα ξεχωριστά θέματα με βάση την συνύπαρξη λέξεων. Για παράδειγμα, στα κείμενα ειδήσεων, οι λέξεις που αντιστοιχούν στην πολιτική, την επιστήμη και τα οικονομικά θα τείνουν να συνυπάρχουν μέσα στο ίδιο έγγραφο. Η μοντελοποίηση του θέματος αντιπροσωπεύει ένα έγγραφο ως σταθμισμένο άθροισμα των θεμάτων, τα οποία μπορούν με τη σειρά τους να χρησιμοποιηθούν ως ενσωματωμένη πρόταση.

Τα NMF και LDA είναι οι δύο δημοφιλέστεροι μέθοδοι για τη μοντελοποίηση θέματος.

4.3.1 Πλεονεκτήματα

Η μοντελοποίηση του θέματος μπορεί εύκολα να ερμηνευτεί και να υπολογιστεί αποτελεσματικά. Η νομιμότητα της μοντελοποίησης μπορεί να εξεταστεί από τα ληφθέντα θέματα. Η μοντελοποίηση του θέματος είναι επίσης εύκολη στην υλοποίηση, με το NMF να υλοποιείται στο sklearn , και το LDA να υλοποιείται με gensim .

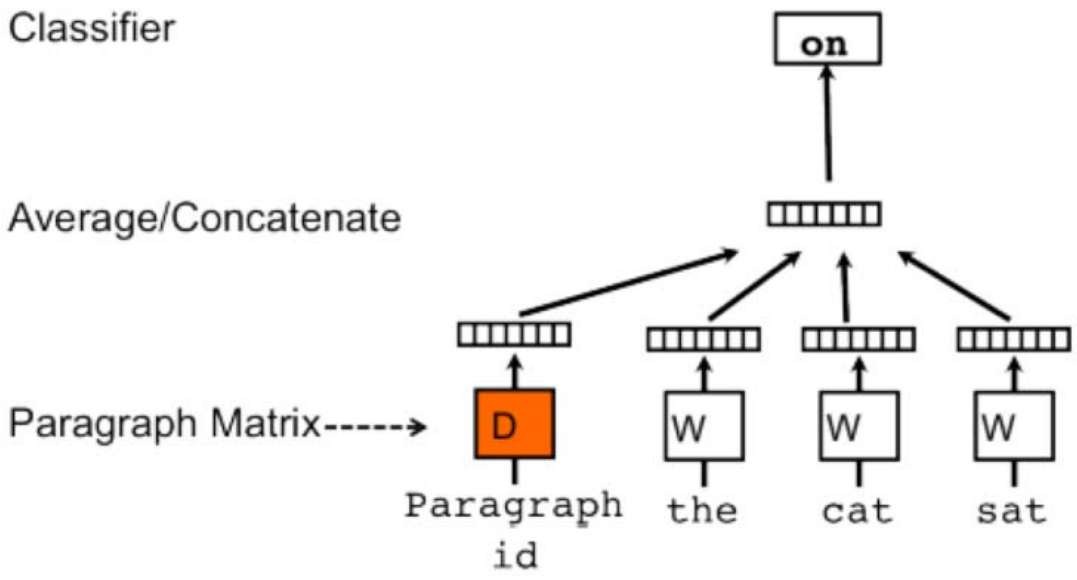
4.3.2 Μειονεκτήματα

Η μοντελοποίηση του θέματος δεν λαμβάνει υπόψη τη σειρά των λέξεων σε μια πρόταση. Δεν καταγράφει ούτε την έννοια της φράσης.

4.4 Διανύσματα παραγράφου (doc2vec)

Εισαγόμενα σε αυτό το έγγραφο , τα διανύσματα παραγράφων (συχνά αναφέρονται ως doc2vec) είναι μια επέκταση του word2vec. Το Word2vec προσπαθεί να προβλέψει μια λέξη (λέξεις) σε μια πρόταση από τις λέξεις που την περιβάλλουν. Οι φορείς παραγράφων επεκτείνουν αυτό τον τρόπο με την κατάρτιση διανυσμάτων παραγράφων ως βοηθητικών πληροφοριών κατά τη διάρκεια αυτής της εργασίας πρόβλεψης.

Οι λεπτομέρειες παρουσιάζονται στο παρακάτω σχήμα.



Εικόνα 4 doc2vec vector[7]

Κάθε παράγραφος (ή πρόταση / έγγραφο) σχετίζεται με ένα διάνυσμα. Κατά τη διάρκεια της κατάρτισης του μοντέλου word2vec, το διάνυσμα παραγράφου είτε βρίσκεται με το μέσο όρο είτε προσκολλάται με το διάνυσμα περιβάλλοντος (που αποτελείται από τους φορείς λέξεων των λέξεων στην πρόταση) και χρησιμοποιείται στην πρόβλεψη.

Διαισθητικά, οι φορείς παραγράφων κωδικοποιούν τις πρόσθετες πληροφορίες που είναι απαραίτητες για την πρόβλεψη μιας λέξης από το περιβάλλον της, η οποία δεν είναι ξεκάθαρη μόνο από τις ίδιες τις λέξεις-πλαίσια. Για παράδειγμα, η ακόλουθη πρόταση

"Έχω ένα κατοικίδιο _____"

[7] https://medium.com/scaleabout/a-gentle-introduction-to-doc2vec-db3e8c0cce5e

θα μπορούσε να περιέχει ένα πλήθος λέξεων, συμπεριλαμβανομένων των "γάτα", "σκύλο", και "φίδι". Ο παράγων παραγράφου θα βοηθούσε την πρόβλεψη με την κωδικοποίηση της λέξης που εισέρχεται στο κενό.

Ως λέξη προσοχής, αυτή είναι μια εξήγηση του μοντέλου κατανεμημένης μνήμης.

4.4.1 Πλεονεκτήματα

Στο πρωτότυπο έγγραφο, οι φορείς παραγράφων εξηγούνται ότι έχουν τα ακόλουθα πλεονεκτήματα:

(1) Κληρονομούν τη σημασιολογία των λέξεων

Για παράδειγμα, η λέξη "power" είναι πιο κοντά στο "ισχυρό" στον σημασιολογικό χώρο σε σύγκριση με το "Paris" (Παρίσι). Οι φορείς παραγράφων μπορούν να χρησιμοποιήσουν αυτές τις μαθησιακές αναπαραστάσεις.

(2) Λαμβάνουν υπόψη τη σειρά λέξεων

Αν και τα διανύσματα παραγράφων δεν χρησιμοποιούν ένα διαδοχικό μοντέλο όπως ένα RNN, λαμβάνουν υπόψη τη σειρά λέξεων σε ένα μικρό πλαίσιο, παρόμοιο με το n-gram μοντέλο με ένα μεγάλο n. Αυτό είναι σημαντικό, αφού τα μοντέλα n-gram διατηρούν πολλές πληροφορίες της παραγράφου, συμπεριλαμβανομένης της σειράς λέξεων. Σε σύγκριση με τα μοντέλα n-gram, οι φορείς παραγράφων είναι σημαντικά μικρότερες διαστάσεις και λιγότερο επιρρεπείς σε υπερφόρτωση όταν χρησιμοποιούνται.

Οι φορείς παραγράφων υλοποιούνται και σε gensim , καθιστώντας τους πολύ εύκολο στη χρήση. Είναι επίσης σχετικά γρήγορο να εκπαιδευτούν.

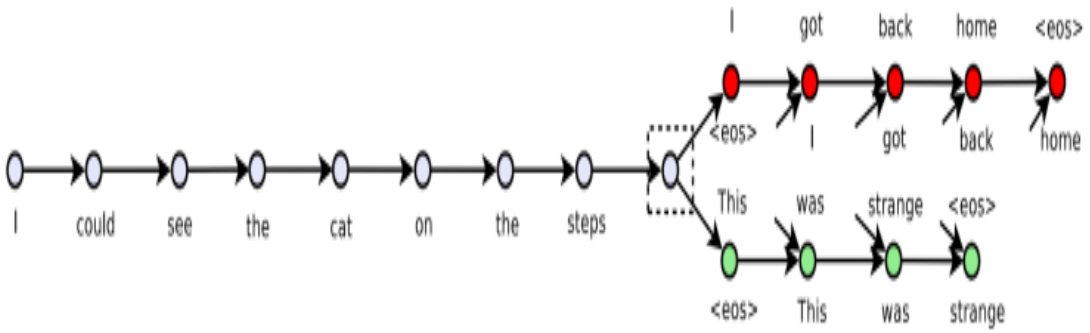
4.4.2 Μειονεκτήματα

Το είδος των πληροφοριών που συλλέγονται στα διανύσματα παραγράφων είναι ασαφές και δύσκολο να ερμηνευτούν, δεδομένου ότι η εργασία πρόβλεψης κωδικοποιεί πληροφορίες τόσο σε διανύσματα λέξεων όσο και σε διανύσματα παραγράφων. Η ποιότητα των διανυσμάτων εξαρτάται επίσης σε μεγάλο βαθμό από την ποιότητα των διανυσμάτων λέξεων.

4.5 Διανύσματα Skip-Thought

Οι φορείς Skip-Thought είναι - με κάποιο τρόπο - word2vec για προτάσεις. Όπου το word2vec προσπαθεί να προβλέψει τα περιβάλλοντα λόγια από συγκεκριμένες λέξεις σε μια πρόταση, ο φορέας Skip-Thought επεκτείνει αυτή την ιδέα στις προτάσεις: προβλέπει τις φράσεις γύρω από μια δεδομένη πρόταση: .

Το μοντέλο συνοψίζεται στο ακόλουθο σχήμα:



Εικόνα 5 skip-thought[8]

[8] https://mlexplained.com/2017/12/28/an-overview-of-sentence-embedding-methods/

Οι φορείς σκέψης Skip-thought χρησιμοποιούν το μοντέλο κωδικοποιητή-αποκωδικοποιητή για να κωδικοποιήσουν πρώτα μια φράση σε ένα διάνυσμα, και στη συνέχεια να αποκωδικοποιήσουν αυτήν την αναπαράσταση στις γύρω προτάσεις.

Οι φορείς σκέψης Skip-Thought έχουν επιτύχει εξαιρετικά αποτελέσματα σε μια ευρεία ποικιλία εργασιών, συμπεριλαμβανομένης της ανάλυσης αισθήσεων και της ανίχνευσης παραφράσεων.

4.5.1 Πλεονεκτήματα

Οι Skip-Thought όχι μόνο λαμβάνουν υπόψη τη σειρά λέξεων, αλλά λαμβάνουν υπόψη και την παραγγελία των προτάσεων . Αυτό τους επιτρέπει να κωδικοποιούν πλούσιες πληροφορίες στην ενσωμάτωση. Δεν αποτελεί έκπληξη το γεγονός ότι οι φορείς που σκέπτονται το skip δείχνουν συχνά την καλύτερη απόδοση μεταξύ άλλων μεθόδων χωρίς επίβλεψη σε ένα ευρύ φάσμα εργασιών.

4.5.2 Μειονεκτήματα

Σε αντίθεση με τις άλλες μεθόδους, οι φορείς που σκέπτονται την παράκαμψη (Skip-Thought) απαιτούν την κατάταξη των προτάσεων με σημασιολογικά σημαντικό τρόπο. Αυτό καθιστά δύσκολη τη χρήση αυτής της μεθόδου για τομείς όπως τα κείμενα στα κοινωνικά μέσα, όπου κάθε απόσπασμα κειμένου υπάρχει ξεχωριστά. Λόγω της ύπαρξης ενός μοντέλου νευρωνικού δικτύου σε βάθος, είναι επίσης πολύ πιο αργά για εκπαίδευση από τις άλλες μεθόδους.

4.6 FastSent

Οι Skip-Thought είναι αργοί . Η FastSent προσπαθεί να αποκαταστήσει αυτήν την αναποτελεσματικότητα, ενώ παράλληλα επεκτείνει την ιδέα του σκεπτικού ότι η

πρόβλεψη των φράσεων γύρω από το περιβάλλον είναι ένας ισχυρός τρόπος για να αποκτήσετε κατανεμημένες αναπαραστάσεις.

Επίσημα, το FastSent αντιπροσωπεύει τις προτάσεις ως το απλό άθροισμα των ενσωματωμένων λέξεων, καθιστώντας την εκπαίδευση αποτελεσματική. Οι "word embeddings" μαθαίνονται έτσι ώστε να μεγιστοποιείται το εσωτερικό προϊόν μεταξύ της ενσωματώσεως της φράσης και της ενσωματωμένης λέξης των φράσεων που περιβάλλουν.

4.6.1 Πλεονεκτήματα

Η FastSent μοιράζεται τα πλεονεκτήματα και τα μειονεκτήματα της Skip-Thought και το είδος των πληροφοριών που κωδικοποιεί. Είναι πολύ πιο αποτελεσματική από το skip-thought, το οποίο είναι και το κύριο πλεονέκτημά της.

4.6.2 Μειονεκτήματα

Η FastSent θυσιάζει τη σειρά λέξεων για λόγους αποτελεσματικότητας, κάτι που μπορεί να είναι ένα μεγάλο μειονέκτημα ανάλογα με τη χρήση.

4.7 Σταθμισμένο άθροισμα των διανυσμάτων λέξεων

Το έγγραφο έδειξε ότι για την ενσωμάτωση προτάσεων μπορεί να λειτουργήσει καλά: ένα σταθμισμένο άθροισμα λέξεων .

Σε αυτή τη μέθοδο, κάθε διάνυσμα λέξης σταθμίζεται από τον παράγοντα \ frac {a} {a + p (w)} όπου a είναι η παράμετρος και p (w) είναι η συχνότητα (εκτιμώμενης) λέξης. Αυτό είναι παρόμοιο με τη στάθμιση tf-idf, όπου οι συχνότεροι όροι σταθμίζονται.

Το έγγραφο δίνει μια ενδιαφέρουσα θεωρητική δικαιολόγηση σε αυτή τη διατύπωση, βασισμένη σε ένα πρότυπο φράσεων όπου κάθε λέξη σε μια πρόταση παράγεται από ένα "τυχαίο περίπατο" του διδακτικού λόγου (sentence).

4.7.1 Πλεονεκτήματα

Λόγω της απλότητας της, αυτή η μέθοδος είναι πολύ εύκολη στην εφαρμογή με το χέρι και γρήγορα για να υπολογιστεί. Είναι επίσης ευρέως εφαρμόσιμο και μπορεί να χρησιμοποιηθεί τόσο για μικρές όσο και για μακριές προτάσεις από μια μεγάλη ποικιλία τομέων συμπεριλαμβανομένων των κοινωνικών μέσων.

4.7.2 Μειονεκτήματα

Δεδομένου ότι πρόκειται για μια σχετικά νέα μέθοδο, το σταθμισμένο άθροισμα των διανυσμάτων λέξεων δεν έχει χρησιμοποιηθεί εκτενώς και δοκιμαστεί σε σύγκριση με παλαιότερες μεθόδους, όπως το skip-thought, καθιστώντας δύσκολη την πρόβλεψη της απόδοσής του. Η σειρά λέξεων και οι φράσεις γύρω από το κείμενο αγνοούνται επίσης, περιορίζοντας τις πληροφορίες που κωδικοποιούνται.

Κεφάλαιο 5

Αλγόριθμος

Παρουσίαση κώδικα

Για την υλοποίηση του κώδικα χρησιμοποιήσαμε python και όλες τις απαραίτητες βιβλιοθήκες όπως φαίνεται και παρακάτω.

Το mainGuardian.py είναι αυτό που τρέχει στους 2 φακέλους των άρθρων για words και το mainGuardianNgrams.py τρέχει με τη λειτουργία των n-grams, όπου εμείς ορίζουμε το μέγεθος τους, τόσο για το φάκελο εκπαίδευσης όσο και για τον φάκελο δοκιμής. Τα προγράμματα τρέχουν σε python .

Προκειμένου να δημιουργηθεί το μοντέλο κατανεμημένης αναπαράστασης, η μέθοδος Doc2vec λαμβάνει ως είσοδο τα έγγραφα απλού κειμένου του σετ εκπαίδευσης χωρίς καμία ετικέτα (δεν υπάρχει καμία περιγραφή της κλάσης στην οποία ανήκουν) και τα μη επισημασμένα έγγραφα κειμένου από το σετ δοκιμών.

Για προ επεξεργασία, πραγματοποιήσαμε μόνο μια τυπική διαδικασία tokenization. Διαφορετικές αναπαραστάσεις των κειμένων χρησιμοποιήθηκαν ως τύποι δεδομένων εισόδου για τη μέθοδο Doc2vec προκειμένου να αξιολογηθεί η ποιότητα των διαφορετικών εξόδων κατανεμημένης αναπαράστασης. Συγκεκριμένα, αναπαραστήσαμε τα κείμενα με όρους λέξεις words και n-grams,για n-grams=2, n-grams=3, n-grams=4 και n-grams=5.

MainGuardian.py

```
from __future__ import division
import codecs
from os import listdir
from os.path import isfile, join, basename
import os
import gensim
import os
import nltk.data
```

```
import numpy
import pathlib
from nltk import ngrams
from nltk.tokenize import word_tokenize
from gensim.models.doc2vec import TaggedDocument, Doc2Vec
from sklearn import svm, metrics
import multiprocessing
cores = multiprocessing.cpu_count()

docLabels = []
folder_name = "C10/test"
test_folder = "C10/train"
authors = []

doc_list = []
dir_list = []
file_list = []
i = 10


def get_ngrams(text,n):
   return [text[i:i+n] for i in range(len(text)-n+1)]


# gia ka8e arxeio poy vrisketai ston fakelo train
# pare tin dieu8unsi tou arxeiou
for dirpath, dirnames, filenames in os.walk(folder_name):
   for name in filenames:
      file_list.append(pathlib.PurePath(dirpath, name))

# gia ka8e arxeio apo tin parapanw lista
```

```
# anoi3e to kai spase to arxeio se protaseis kai topo8etisetes se lista
for file in file_list:
    f1 = codecs.open(str(file), "r", "utf-8", errors='ignore')
    st = f1.read()

    authors.append(basename(os.path.abspath(os.path.join(str(file), os.pardir))))

    sent_text = nltk.sent_tokenize(st)
    for sentence in sent_text:
        doc_list.append(sentence)
        docLabels.append(i)
        i=i+1

taggeddoc=[]

for doc, index in zip(doc_list, docLabels):
    #print(nltk.word_tokenize(doc))
    td = TaggedDocument(words=nltk.word_tokenize(doc), tags=[str(index)])
    #td = TaggedDocument(words=get_ngrams(doc,4), tags=[str(index)])
    taggeddoc.append(td)

#dimiourgia montelou apo to sunolo protasewn twn keimenwn
model = gensim.models.Doc2Vec(taggeddoc,  dm = 0,window=2, alpha=0.025, size= 100,
min_alpha=0.050, min_count=2, workers =cores)

for epoch in range(10):
    if epoch % 2 == 0:
        print ('Now training epoch %s'%epoch)
    model.train(taggeddoc, total_examples=model.corpus_count, epochs=model.iter)
    model.alpha -= 0.002  # decrease the learning rate
    model.min_alpha = model.alpha  # fix the learning rate, no decay
```

```
    model.train(taggeddoc, total_examples=model.corpus_count, epochs=model.iter)


model.save("doc2vec.model")

model = Doc2Vec.load("doc2vec.model")

all_vectors=[]

for file in file_list:
    f1 = codecs.open(str(file), "r", "utf-8", errors='ignore')
    st = f1.read()
    #xorizoume to keimeno se protaseis
    sents = nltk.sent_tokenize(st)

    sentences = []

    for sen in sents:
        doc =  get_ngrams(sen, 3)
        sentences.extend(doc)
        new_vector = model.infer_vector(sentences)
    all_vectors.append(new_vector)




kernels = ['linear', 'rbf', 'poly']
cs = [0.1, 1, 10, 100, 1000]
gammas = [0.1, 1, 10, 100]
```

```
svc = svm.SVC(kernel='linear', C=1000, gamma=10)
svc.fit(all_vectors,authors)

print("----------------------------")



list_of_files =0
file_list = []
#gia ka8e arxeio poy vrisketai ston fakelo train pare tin dieu8unsi tou arxeiou
for dirpath, dirnames, filenames in os.walk(test_folder):
   for name in filenames:
      #metraei to plithos twn arxeiwn sto test
      list_of_files += 1
      file_list.append(pathlib.PurePath(dirpath, name))



all_vectors=[]
authors = []
for file in file_list:
   f1 = codecs.open(str(file), "r", "utf-8", errors='ignore')
   st = f1.read()
   authors.append(basename(os.path.abspath(os.path.join(str(file), os.pardir))))
   #xorizoume to keimeno se protaseis
   sents = nltk.sent_tokenize(st)



   sentences = []
   for sen in sents:
      #xwrizoume to keimeno se lexeis
      doc =  get_ngrams(sen, 1)
      sentences.extend(doc)
```

```
    new_vector = model.infer_vector(sentences)
    #print (sentences)
    all_vectors.append(new_vector)



print ("authors" + str(len(authors)))
#print ("All " + str(len(all_vectors)))
predicted = svc.predict(all_vectors)
print ("predicted" + str(len(predicted)))



list_common = []
for a, b in zip(authors, predicted):
    if a == b:
        list_common.append(a)

print (len(list_common))
print (list_common)
print ("Swsta vre8ikan "+str(len(list_common))+" apo ta "+str(len(authors)))

pososto = len(list_common)/len(authors)

print ("epitixia : " , pososto*100,"%")
```

Κεφάλαιο 6

Αποτελέσματα

Πραγματοποιήσαμε πειράματα πάνω από τα διαφορετικά σύνολα δεδομένων που περιγράφονται στην προηγούμενη ενότητα, χρησιμοποιώντας μια προσέγγιση μηχανικής μάθησης.

Τα πειράματα περιελάμβαναν στην εκπαίδευση ενός ταξινομητή με το αντίστοιχο σύνολο εκπαίδευσης χρησιμοποιώντας την κατανεμημένη αναπαράσταση εγγράφων και στη συνέχεια την αξιολόγηση του ταξινομητή με το σετ δοκιμών χρησιμοποιώντας την ίδια αναπαράσταση.

Αναφέρουμε τα αποτελέσματά μας ως προς την ακρίβεια που λαμβάνουμε για κάθε σύνολο δεδομένων με τη χρήση λέξεων(words) και n-grams (n = 2,3,4,5) ως τύπου δεδομένων εισόδου για τη μέθοδο Doc2vec, προκειμένου να πάρουμε την κατανεμημένη αναπαράσταση των εγγράφων.

Η ακρίβεια αντιπροσωπεύει το ποσοστό των στιγμιότυπων που έχουν ταξινομηθεί σωστά και ορίζεται ως πραγματικές θετικές τιμές (TP), ψευδώς θετικά (FP), αληθή αρνητικά (TN) και ψευδώς αρνητικά (FN) ως εξής:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Χρησιμοποιήσαμε τη μέθοδο Doc2vec (Le και Mikolov 2014) που είναι διαθέσιμη ελεύθερα στη μονάδα GENSIM.

Το υποπρόγραμμα Doc2vec χρησιμοποιεί αλγόριθμους μετωπικής κλίσης και αλγόριθμους οπτικής διάδοσης για τη δημιουργία του μοντέλου, ωστόσο, αυτοί οι αλγόριθμοι χρησιμοποιούν τυχαία στοιχεία που δεν εγγυώνται την αναπαραγωγικότητά τους.

Προκειμένου να διασφαλιστεί η αναπαραγωγικότητα των πειραμάτων μας, οι τιμές των ακόλουθων παραμέτρων είναι σταθερές (σύμφωνα με τις συστάσεις του εγχειριδίου χρήσης).Η τιμή του κατώτατου ορίου για τη διαμόρφωση των λέξεων υψηλότερης συχνότητας είναι τυχαία υποδεικνυόμενη 1e-3, η αρνητική δειγματοληψία ορίζεται στο 5, η τροφοδοσία της γεννήτριας τυχαίων αριθμών έχει οριστεί σε 1 και ο

αριθμός των νημάτων έχει οριστεί σε core (δεν χρησιμοποιείται multithreading). Οι υπόλοιπες παράμετροι έχουν οριστεί με τις προεπιλεγμένες τιμές.

Στην παρούσα εργασία παρουσιάζουμε μια προσέγγιση που χρησιμοποιεί τη μέθοδο Doc2vec για την εκμάθηση της αναπαράστασης κατανεμημένων εγγράφων και την εποπτευόμενη μέθοδο μηχανικής μάθησης για την εφαρμογή της ανάθεσης συγγραμμάτων. Εκτός από τη χρήση λέξεων ως τυπικού τύπου δεδομένων εισόδου για την κατασκευή των κατανεμημένων αναπαραστάσεων, προτείνουμε έναν νέο τύπο δεδομένων εισόδου βασισμένο στις λέξεις (χαρακτήρες)n-grams (n = 1, 2, 3, 4, 5). Η συμπερίληψη των χαρακτήρων n-grams μας επιτρέπει να αποκτήσουμε περαιτέρω γνώση της αποτελεσματικότητας του μοντέλου για να μαθαίνει τα συντακτικά και γραμματικά πρότυπα ενός συγγραφέα.

Τα πειράματα[16] έδειχναν ότι η δημιουργία μοντέλων κατανεμημένης αναπαράστασης κάθε προτεινόμενου τύπου δεδομένων εισόδου (λέξεων και n-grams με n = 2, 3, 4, 5), επιτυγχάνει τα καλύτερα αποτελέσματα σε διαφορετικά σύνολα δεδομένων.

Τα σύνολα δεδομένων που χρησιμοποιήθηκαν στα πειράματά μας πρόσφεραν μια ποικιλία σεναρίων για να δοκιμάσουν την πρότασή μας και το πιο ενδιαφέρον ήταν το Guardian10 επειδή μας έδωσε τη δυνατότητα να αξιολογήσουμε την πρότασή μας πάνω σε κείμενα δύο τομέων και διασταυρούμενου είδους με μικρό μήκος.

Παρατηρήθηκε ότι τα καλύτερα αποτελέσματα τα παίρνουμε όταν γίνετε η χρήση των κείμενων ως κάθε αυτό λέξεων και πουθενά η χρήση των χαρακτήρων ngrams δεν μας έχει δώσει καλύτερα αποτελέσματα, σε αντίθεση με το paper των Posadas Duran et al[16] όπου η χρήση words **και** n-grams έδινε πάντα τα καλύτερα αποτελέσματα. Οι κατανεμημένες αναπαραστάσεις των 2 κειμένων ελέγχονται για ταυτίσεις και το ποσοστό που βγαίνει είναι βάση αυτής της σύγκρισης.

Ουσιαστικά η διαφοροποίηση μας στην υλοποίηση των πειραμάτων μας με αυτά των Posadas Duran et al[16] είναι ότι εμείς δοκιμάσαμε την μέθοδο μας συγκρίνοντας τα αποτελέσματα που μας έδιναν τα ξεχωριστά για κάθε ένα σύνολο δεδομένων που παράγαμε είτε για στοιχεία εισόδου ως λέξεις(words) είτε ως n-grams ξεχωριστά κάθε φορά (για n= 2, 3, 4, 5). Με αυτό τον τρόπο επιχειρήσαμε να δούμε τις ομοιότητες συγγραφής μεταξύ διαφορετικών θεματικών κειμένων των ίδιων συγγραφέων λέξη προς λέξη όπου εκεί είχαμε και την μεγαλύτερη επιτυχία στα

πειράματά μας, καθώς και με τη χρήση χαρακτήρων n-grams, χωρίζοντας τα κείμενα σε χαρακτήρες 2-grams, 3-grams, 4-grams, και 5-grams κάθε φορά ξεχωριστά, συγκρίνοντας πάλι έτσι τις ομοιότητες μεταξύ των κειμένων διαφορετικής θεματολογίας, από ένα κλειστό γκρουπ "γνωστών" συγγραφέων του corpus Guardian10.

Στους παρακάτω πίνακες βλέπουμε τις συγκριτικές τιμές των πειραμάτων μας (με πράσινο χρώμα) SVM Ktis με αυτές των πειραμάτων των Posadas Duran et al (με κίτρινο χρώμα) LR-PD.

Γίνετε αναφορά στον αλγόριθμο ταξινόμησης που χρησιμοποιήθηκε για αυτά τα αποτελέσματα καθώς όπως αναφέρουν οι Posadas Duran et al στην εργασία τους[16], τα αποτελέσματα που έχουν πάρει έχουν προκύψει από τη χρήση του LR(Logistic Regression) ταξινομητή, παρόλο που είχαν πάρει αποτελέσματα και με SVM, χρησιμοποιήθηκαν αυτά του LR καθώς έδινε καλύτερα αποτελέσματα. Αντίθετα τα δικά μας αποτελέσματα είναι με SVM και παρά το ότι χρησιμοποιήθηκαν οι ίδιες τιμές παραμετροποίησης για το doc2vec των πειραμάτων μας μέσα στον αλγόριθμο μας, με αυτές των Posadas Duran et al στα δικά τους πειράματα αντίστοιχα, οι τιμές που πήραμε εμείς έχουν αισθητή διαφορά από τις τιμές των Posadas Duran et al.

Books	Politics LR-PD	SVM-Ktis	Society LR-PD	SVM-Ktis	UK LR-PD	SVM-Ktis	World LR-PD	SVM-Ktis
Words	56,25%	31,25%	51,61%	20,97%	62,22%	37,78%	51,28%	41,03%
ngrams2	60,71%	12,50%	59,68%	9,68%	62,22%	16,67%	48,72%	15,56%
ngrams3	58,93%	20,54%	58,06%	17,74%	60,00%	16,67%	47,01%	18,89%
ngrams4	58,93%	17,86%	58,06%	14,52%	60,00%	24,44%	47,01%	20,00%
ngrams5	58,93%	10,71%	58,06%	9,68%	60,00%	20,00%	47,01%	16,67%

Πίνακας 7 συγκριτικών ποσοστών των πειραματικών αποτελεσμάτων του αλγορίθμου μας με αυτά των Posadas-Duran et al[16]. Εδώ για το Guardian10 Corpus με training το σύνολο Books και για test όλα τα υπόλοιπα σύνολα.

Politics	Books LR-PD	SVM-Ktis	Society LR-PD	SVM-Ktis	UK LR-PD	SVM-Ktis	World LR-PD	SVM-Ktis
Words	87,30%	44,44%	95,16%	48,39%	91,11%	60,00%	93,16%	52,14%
ngrams2	93,65%	15,87%	98,39%	29,03%	95,56%	24,19%	91,45%	17,94%
ngrams3	92,06%	25,40%	98,39%	37,10%	93,33%	35,48%	91,45%	32,48%
ngrams4	93,65%	31,75%	98,39%	24,19%	93,33%	29,03%	91,45%	22,22%
ngrams5	93,65%	23,81%	98,39%	30,65%	93,33%	29,03%	91,45%	31,62%

Πίνακας 8 συγκριτικών ποσοστών των πειραματικών αποτελεσμάτων του αλγορίθμου μας με αυτά των Posadas-Duran et al[16]. Εδώ για το Guardian10 Corpus με training το σύνολο Politics και για test όλα τα υπόλοιπα σύνολα.

Society	Books LR-PD	SVM-Ktis	Politics LR-PD	SVM-Ktis	UK LR-PD	SVM-Ktis	World LR-PD	SVM-Ktis
Words	57,14%	22,22%	64,29%	32,14%	55,56%	38,39%	60,68%	36,75%
ngrams2	53,97%	14,29%	65,18%	17,86%	61,11%	17,78%	63,25%	16,24%
ngrams3	53,97%	11,11%	65,18%	25,00%	60,00%	20,00%	63,25%	25,64%
ngrams4	53,97%	12,70%	64,29%	21,43%	60,00%	20,00%	63,25%	27,35%
ngrams5	53,97%	17,46%	64,29%	27,68%	60,00%	15,56%	63,25%	24,79%

Πίνακας 9 συγκριτικών ποσοστών των πειραματικών αποτελεσμάτων του αλγορίθμου μας με αυτά των Posadas-Duran et al[16]. Εδώ για το Guardian10 Corpus με training το σύνολο Society και για test όλα τα υπόλοιπα σύνολα.

UK	Books LR-PD	SVM-Ktis	Politics LR-PD	SVM-Ktis	Society LR-PD	SVM-Ktis	World LR-PD	SVM-Ktis
Words	82,54%	44,44%	78,57%	41,96%	90,32%	45,16%	78,63%	40,17%
ngrams2	88,89%	14,29%	87,50%	22,32%	93,55%	27,42%	83,76%	10,26%
ngrams3	90,48%	20,63%	87,50%	23,21%	95,16%	35,48%	82,91%	27,35%
ngrams4	88,89%	23,81%	87,50%	27,68%	95,16%	25,81%	82,91%	20,51%
ngrams5	88,89%	23,81%	87,50%	26,79%	95,16%	17,74%	82,91%	11,97%

Πίνακας 10 συγκριτικών ποσοστών των πειραματικών αποτελεσμάτων του αλγορίθμου μας με αυτά των Posadas-Duran et al[16]. Εδώ για το Guardian10 Corpus με training το σύνολο UK και για test όλα τα υπόλοιπα σύνολα.

World	Books LR-PD	SVM-Ktis	Politics LR-PD	SVM-Ktis	Society LR-PD	SVM-Ktis	UK LR-PD	SVM-Ktis
Words	77,78%	31,75%	78,57%	43,75%	87,10%	41,94%	80,00%	43,33%
ngrams2	77,78%	17,46%	83,93%	20,54%	93,55%	12,90%	83,33%	14,44%
ngrams3	74,60%	22,22%	86,61%	24,11%	93,55%	30,65%	83,33%	21,11%
ngrams4	74,60%	17,46%	85,71%	21,43%	93,55%	30,65%	82,22%	12,22%
ngrams5	74,60%	19,05%	85,71%	23,21%	93,55%	27,42%	82,22%	15,56%

Πίνακας 11 συγκριτικών ποσοστών των πειραματικών αποτελεσμάτων του αλγορίθμου μας με αυτά των Posadas-Duran et al[16]. Εδώ για το Guardian10 Corpus με training το σύνολο World και για test όλα τα υπόλοιπα σύνολα.

(Έχουν χρησιμοποιηθεί οι ίδιες τιμές παραμετροποίησης με αυτές των Posadas-Duran et al[16] για την εκπαίδευση του doc2vec μέσα στον αλγόριθμο).

Κεφάλαιο 7

Συμπεράσματα

Στην παρούσα εργασία παρουσιάζουμε μια προσέγγιση που χρησιμοποιεί τη μέθοδο Doc2vec για την εκμάθηση της αναπαράστασης κατανεμημένων εγγράφων και εποπτευόμενη μέθοδο μηχανικής μάθησης για την εφαρμογή της ανάθεσης συγγραμμάτων. Προσπαθώντας να προσεγγίσουμε την μέθοδο που ακολούθησαν οι Posadas Duran et al και παρουσίασαν στο paper τους με τίτλο "Application of the distributed document representation in the authorship attribution task for small corpora" .

Εκτός από τη χρήση λέξεων ως τυπικού τύπου δεδομένων εισόδου για την κατασκευή των κατανεμημένων αναπαραστάσεων, προτάθηκε και ένας τύπος δεδομένων εισόδου βασισμένο στους χαρακτήρες n-grams (n = 2, 3, 4, 5). Η χρήση χαρακτήρων n-grams μας επιτρέπει να αποκτήσουμε γνώση της αποτελεσματικότητας του μοντέλου για να μαθαίνει τα συντακτικά και γραμματικά πρότυπα ενός συγγραφέα.

Τα πειράματα που διεξήγαγαν οι Posadas Duran et al σε διαφορετικά σύνολα δεδομένων έδειξαν ότι η χρήση της μεθόδου Doc2vec για την εκμάθηση αναπαραστάσεων κατανεμημένων εγγράφων, στις περισσότερες περιπτώσεις, υπερέχει των αποτελεσμάτων προηγούμενων εργασιών. Εμείς παρόλα αυτά δεν είδαμε κάτι τέτοιο στη δική μας πειραματική διαδικασία. Επίσης τα πειράματα των Posadas Duran et al έδειξαν ότι η δημιουργία μοντέλων κατανεμημένης αναπαράστασης κάθε προτεινόμενου τύπου δεδομένων εισόδου (λέξεων n-grams με n = 2, 3, 4, 5) ανεξάρτητων και στη συνέχεια με συνδυασμό τους σε ένα μόνο φορέα επιτυγχάνει τα καλύτερα αποτελέσματα (τις περισσότερες φορές) σε διαφορετικά σύνολα δεδομένων.

Η αναπαράσταση χρησιμοποιώντας λέξεις και μεγάλα grams λέξεων ήταν ο πιο ενημερωτικός συνδυασμός από άλλους προτεινόμενους τύπους δεδομένων εισόδου. Σε εμάς αντίθετα συγκρίνοντας μόνο τα ngrams που παράγαμε από τα σύνολα δεδομένων δεν είδαμε βελτιωμένα αποτελέσματα, κάθε άλλο τα καλύτερα αποτελέσματα τα πήραμε όταν έγινε χρήση ολόκληρων των λέξεων(words) των κειμένων.

Εικόνες/ Πίνακες

Σχήμα 1 Το Corpus Guardian για την κλειστή καταχώριση συγγραφέων, με το πολύ 10 έγγραφα ανά συγγραφέα.[16] 20
Εικόνα 2 Αρχιτεκτονική του doc2vec https://www.slideshare.net/ibelmopan/text-mining-lab-summer-2017-word-vector-representation 27
Εικόνα 3 Paragraph 2 vec https://www.slideshare.net/BhaskarMitra3/neural-text-embeddings-for-information-retrieval-wsdm-2017 29
Εικόνα 4 doc2vec vector https://medium.com/scaleabout/a-gentle-introduction-to-doc2vec-db3e8c0cce5e 40
Εικόνα 5 skip-thought https://mlexplained.com/2017/12/28/an-overview-of-sentence-embedding-methods/ 42

Πίνακας 1 Αποτελέσματα του corpus Guardian υπολογίζοντας ως δεδομένα εκπαίδευσης των φάκελο **politics** και ως δεδομένα τεστ όλους τους υπόλοιπους φάκελους δεδομένων των Posadas Duran et al16] 17
Πίνακας 2 Αποτελέσματα του corpus Guardian υπολογίζοντας ως δεδομένα εκπαίδευσης των φάκελο **society** και ως δεδομένα τεστ όλους τους υπόλοιπους φάκελους δεδομένων των Posadas Duran et al[16] 19
Πίνακας 3 Αποτελέσματα του corpus Guardian υπολογίζοντας ως δεδομένα εκπαίδευσης των φάκελο **UK** και ως δεδομένα τεστ όλους τους υπόλοιπους φάκελους δεδομένων των Posadas Duran et al[16] 19
Πίνακας 4 Αποτελέσματα του corpus Guardian υπολογίζοντας ως δεδομένα εκπαίδευσης των φάκελο **World** και ως δεδομένα τεστ όλους τους υπόλοιπους φάκελους δεδομένων των Posadas Duran et al[16] 19
Πίνακας 5 Αποτελέσματα του corpus Guardian υπολογίζοντας ως δεδομένα εκπαίδευσης των φάκελο **Book reviews** και ως δεδομένα τεστ όλους τους υπόλοιπους φάκελους δεδομένων των Posadas Duran et al[16] 19
Πίνακας 6 Αποτελέσματα του corpus Guardian μέσω όρο καλύτερων αποτελεσμάτων για κάθε φάκελο δεδομένων των Posadas Duran et al[16] 20
Πίνακας 7 Συγκριτικών ποσοστών των πειραματικών αποτελεσμάτων του αλγορίθμου μας με αυτά των Posadas-Duran et al[16]. Εδώ για το Guardian10 Corpus με training το σύνολο **Books** και για test όλα τα υπόλοιπα σύνολα. 55
Πίνακας 8 Συγκριτικών ποσοστών των πειραματικών αποτελεσμάτων του αλγορίθμου μας με αυτά των Posadas-Duran et al[16]. Εδώ για το Guardian10 Corpus με training το σύνολο **Politics** και για test όλα τα υπόλοιπα σύνολα. 55
Πίνακας 9 Συγκριτικών ποσοστών των πειραματικών αποτελεσμάτων του αλγορίθμου μας με αυτά των Posadas-Duran[16]. Εδώ για το Guardian10 Corpus με training το σύνολο **Society** και για test όλα τα υπόλοιπα σύνολα. 55
Πίνακας 10 Συγκριτικών ποσοστών των πειραματικών αποτελεσμάτων του αλγορίθμου μας με αυτά των Posadas-Duran et al[16]. Εδώ για το Guardian10 Corpus με training το σύνολο **UK** και για test όλα τα υπόλοιπα σύνολα. 55
Πίνακας 11 Συγκριτικών ποσοστών των πειραματικών αποτελεσμάτων του αλγορίθμου μας με αυτά των Posadas-Duran et al[16]. Εδώ για το Guardian10 Corpus με training το σύνολο **World** και για test όλα τα υπόλοιπα σύνολα. 55

Βιβλιογραφία

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean. "Distributed Representations of Words and Phrases and their Compositionality". part of Advances in Neural Information Processing Systems 26 (NIPS 2013) arXiv:1310.4546 Freely accessible [cs.CL].

[2] Lebret Rémi, Collobert Ronan. "Word Embeddings through Hellinger PCA". Conference of the European Chapter of the Association for Computational Linguistics (EACL). (2014). arXiv:1312.5542 Freely accessible. Bibcode:2013arXiv1312.5542L.

[3] Levy Omer, Goldberg Yoav. Neural Word Embedding as Implicit Matrix Factorization(PDF) NIPS (2014).

[4] Li, Yitan, Xu, Linli Word Embedding Revisited: A New Representation Learning and Explicit Matrix Factorization Perspective (PDF). Int'l J. Conf. on Artificial Intelligence (IJCAI). (2015).

[5] Globerson Amir "Euclidean Embedding of Co-occurrence Data" (PDF). Journal of Machine learning research. (2007).

[6] Qureshi, M. Atif; Greene, Derek ."EVE: explainable vector based embedding technique using Wikipedia". Journal of Intelligent Information Systems. (2018-06-04) arXiv:1702.06891 Freely accessible. doi:10.1007/s10844-018-0511-x. ISSN 0925-9902.

[7] Levy Omer, Goldberg Yoav. Linguistic Regularities in Sparse and Explicit Word(PDF). Publisher: Association for Computational Linguistics Representations. Volume: Proceedings of the Eighteenth Conference on Computational Natural Language Learning (2014) CoNLL. pp. 171–180.

[8] Socher Richard, Bauer John, Manning Christopher, Ng Andrew .Parsing with compositional vector grammars (PDF). Publisher: Association for Computational Linguistics Volume: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2013). Proc. ACL Conf.

[9] Socher Richard, Perelygin Alex, Wu Jean, Chuang Jason, Manning Chris, Ng Andrew, Potts Chris . Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank (PDF).(2013) EMNLP.

[10] Firth J.R. "A synopsis of linguistic theory 1930-1955". Studies in Linguistic Analysis. Oxford: Philological Society: 1–32. Reprinted in F.R. Palmer, ed. (1968). Selected Papers of J.R. Firth 1952-1959. London: Longman. (1957)

[11] Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, Jean-Luc Gauvain "A Neural Probabilistic Language Model". Studies in Fuzziness and Soft Computing: 137–186. doi:10.1007/3-540-33486-6_6.

[12] Lavelli Alberto, Sebastiani Fabrizio, Zanoli Roberto. Distributional term representations: an experimental comparison. 13th ACM International Conference on Information and Knowledge Management. pp. 615–624. (2004)

[13] Roweis Sam T., Saul Lawrence K. "Nonlinear Dimensionality Reduction by Locally Linear Embedding". Science. 290 (5500): 2323. Bibcode:2000Sci...290.2323R. doi:10.1126/science.290.5500.2323. PMID 11125150. (2000)

[14] Andriy Mnih, Geoffrey E. Hinton "A Scalable Hierarchical Distributed Language Model". Advances in Neural Information Processing Systems 21 (NIPS 2008)

[15] Camacho-Collados, Jose, Pilehvar Mohammad Taher. From Word to Sense Embeddings: A Survey on Vector Representations of Meaning. Published in J. Artif. Intell. Res. 2018 DOI:10.1613/jair.1.11259 arXiv:1805.04032 Freely accessible. Bibcode:2018arXiv180504032C. (2018)

[16] Juan-Pablo Posadas-Durán, Helena Gómez-Adorno, Grigori Sidorov, Ildar Batyrshin, David Pinto, Liliana Chanona-Hernández. 'Application of the distributed document representation in the authorship attribution task for small corpora' Springer-Verlag Berlin Heidelberg (2016)