MAI4CAREU

Natural Language Processing

Vector Semantics

Demetris Paschalides

Department of Computer Science

University of Cyprus

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Υπολογιστικά Μοντέλα Σημασιολογίας Λέξεων

- Μπορούμε να οικοδομήσουμε μια θεωρία για το πώς να αναπαραστήσουμε το νόημα των λέξεων, που αντιπροσωπεύει τουλάχιστον ένα μέρος της λεξικής σημασιολογίας;
- Θα παρουσιάσουμε τη Διανυσματική Σημασιολογία:
  - Τυποποιημένο μοντέλο γλωσσικής επεξεργασίας. Χρησιμοποιείται σε διάφορες εργασίες NLP.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Υπόθεση κατανομής

- “Η έννοια μιας λέξης είναι η χρήση της στη γλώσσα”
  - *Wittgenstein PI #43*

- “Θα ξέρετε μια λέξη από την παρέα της”
  - *Firth 1957*

- “Εάν το A και το B έχουν σχεδόν πανομοιότυπα περιβάλλοντα λέμε ότι είναι συνώνυμα”
  - *Harris 1954*

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα: Τι σημαίνει το OngChoi;

- Ας υποθέσουμε ότι βλέπετε τις ακόλουθες προτάσεις:
  - Ongchoi είναι πεντανόστιμο σοταρισμένο με σκόρδο
  - Ongchoi είναι υπέροχο πάνω από το ρύζι
  - Ongchoi φύλλα με αλμυρές σάλτσες

- Επίσης, αυτά:
  - … σπανάκι σοταρισμένο με σκόρδο πάνω από ρύζι
  - Chard stems και τα φύλλα είναι νόστιμα
  - Collard greens και άλλα αλμυρά φυλλώδη χόρτα

→ Το Ongchoi είναι ένα φυλλώδες πράσινο όπως το σπανάκι, τα Chard ή τα χόρτα Collard

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Υπολογιστικές Προσεγγίσεις για Σημασία λέξης

1. Ορισμός νοήματος με γλωσσική κατανομή
   - Ορίστε τη σημασία μιας λέξης με τη διανομή της στη χρήση της γλώσσας, δηλαδ
     τις γειτονικές της λέξεις ή γραμματικά περιβάλλοντα.

2. Σημασία ως σημείο στο διάστημα (*Osgood et al*. 1957):
   - 3 συναισθηματικές διαστάσεις για μια λέξη: Σθένος, Διέγερση, Κυριαρχία
     Η χροιά μιας λέξης είναι ένα διάνυσμα σε 3-διάστατο χώρο.

	Word	Score		Word	Score
Valence	love	1.000		toxic	0.008
	happy	1.000		nightmare	0.005
Arousal	elated	0.960		mellow	0.069
	frenzy	0.965		napping	0.046
Dominance	powerful	0.991		weak	0.045
	leadership	0.983		empty	0.081

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Συνδυάστε τις προσεγγίσεις 1 και 2

- Ορισμός της σημασίας λέξης ως σημείου στο διάστημα, με βάση την κατανομή λέξεων.
- Κάθε λέξη αναπαρίσταται ως διάνυσμα.
  - Όχι μόνο “good” ή w_i
- Παρόμοιες λέξεις βρίσκονται κοντά σε ένα σημασιολογικό χώρο.
  Δημιουργήστε αυτόν τον χώρο αυτόματα, κοιτάζοντας λέξεις που βρίσκονται κοντά.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Ορισμός νοήματος ως διανύσματος

- Αυτά τα διανύσματα ονομάζονται "Ενσωματώσεις" επειδή είναι ενσωματωμένα σε ένα χώρο (π.χ. σημασιολογικός χώρος).
- Οι ενσωματώσεις τυποποιούνται ως ο τρόπος αναπαράστασης του νοήματος στην Επεξεργασία Φυσικής Γλώσσας.
  - Κάθε σύγχρονος αλγόριθμος NLP χρησιμοποιεί ενσωματώσεις ως αναπαράσταση της σημασίας των λέξεων. Είναι ένα λεπτόκοκκο μοντέλο νοήματος για ομοιότητα.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Διαίσθηση: Γιατί Διανύσματα;

- Εστω το task του Sentiment Analysis:
  - Με τις λέξεις, ένα χαρακτηριστικό είναι μια ταυτότητα λέξης:
    - Η προηγούμενη λέξη ήταν w_5 = *terrible*
    - Απαιτεί την ίδια ακριβώς λέξη στην εκπαίδευση και τη δοκιμή.
  - Με embeddings:
    - Χαρακτηριστικά είναι ένα διάνυσμα
    - Η προηγούμενη λέξη ήταν [35, 22, 17, …]
    - Στο σύνολο δοκιμής ένα παρόμοιο διάνυσμα [34, 21, 14, …]
    - Μπορούμε να γενικεύσουμε σε παρόμοιες αλλά αθέατες λέξεις!

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Αναπαράσταση λέξεων
Term-Document Matrix

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	17
soldier	2	80	62	89
fool	36	58	1	4
clown	20	15	2	3

- Context = που εμφανίζεται στο ίδιο έγγραφο.
- **Vector Space Model**: Κάθε έγγραφο αναπαρίσταται ως διάνυσμα στήλης διαστάσεων V.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Αναπαραστάσεις βάσει καταμέτρησης

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

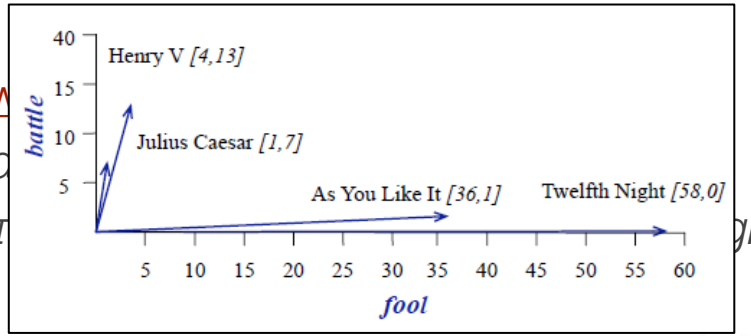
- **Counts = term-frequency**
- Αναπαράσταση λέξης ως διανύσματος (each row)
  - battle είναι "*το είδος της λέξης που εμφανίζεται στο Julius Caesar and Henry V*"
  - fool είναι "*το είδος των λέξεων που εμφανίζονται στις κωμωδίες, ειδικά Twelfth Night*"

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Αναπαραστάσεις βάσει καταμέτρησης

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

- **Counts = term-frequency**
- Αναπαράσταση λέξης ως διανύσματος (each row
  - battle είναι "*το είδος της λέξης που εμφανίζετα*
  - fool είναι "*το είδος των λέξεων που εμφανίζοντ*ght"



Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Υπολογιστική ομοιότητα λέξεων

- Το γινόμενο κουκκίδας (εσωτερικό) μεταξύ δύο διανυσμάτων είναι ένας αριθμός:

$$\text{dot product}(\mathbf{v},\mathbf{w}) = \mathbf{v}\cdot\mathbf{w} = \sum_{i=1}^{N} v_i w_i = v_1 w_1 + v_2 w_2 + \ldots + v_N w_N$$

- Τείνει να είναι υψηλή όταν δύο διανύσματα έχουν μεγάλες τιμές στις ίδιες διαστάσεις.
- Το γινόμενο κουκκίδας μπορεί να είναι χρήσιμη μέτρηση ομοιότητας μεταξύ διανυσμάτων.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Raw Dot-product Issues

- Το γινόμενο ευνοεί τα μακρά διανύσματα.
  - Είναι υψηλότερο εάν ένα διάνυσμα είναι μεγαλύτερο (έχει υψηλότερες τιμές σε πολλές διαστάσεις)
  - Μήκος διανύσματος: $$|\mathbf{v}| = \sqrt{\sum_{i=1}^{N} v_i^2}$$
- Συχνές λέξεις ("*of*", "*the*", "*you*") έχουν μακρά διανύσματα
  - Το συμβαίνει πολλές φορές με άλλες λέξεις.

→ **Το γινόμενο ευνοεί υπερβολικά τις συχνές λέξεις.**

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Συνημίτονο ως μέτρηση ομοιότητας

$$\text{cosine}(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

Με βάση τον ορισμό του γινομένου κουκκίδας μεταξύ των διανυσμάτων a και b:

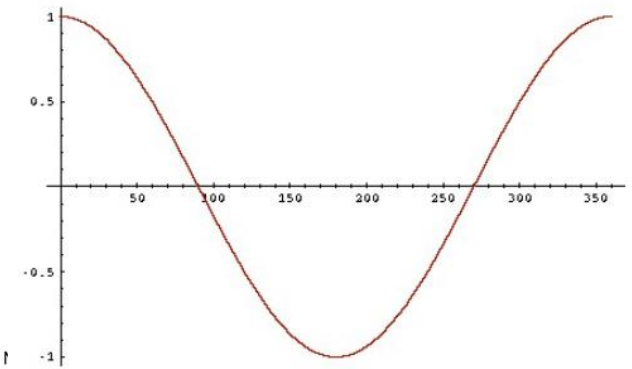
$$\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta$$

$$\frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}||\mathbf{b}|} = \cos\theta$$

- ❑ - 1 για διαν. που δείχνουν σε αντίθετες κατευθύνσεις
- ❑ +1 για διαν. που δείχνουν προς την ίδια κατεύθυνση

  0 για διανύσματα που είναι ορθογώνια

Τιμές συχνότητας >= 0 → συνημίτονο από 0 - 1



Co-financed by the European Union
Connecting Europe Facility
University of Cyprus

Cosine Examples

$$\cos(\vec{v},\vec{w}) = \frac{\vec{v}\bullet\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\vec{v}}{|\vec{v}|}\bullet\frac{\vec{w}}{|\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

	pie	data	computer
cherry	442	8	2
digital	5	1683	1670
information	5	3982	3325

$$\cos(\text{cherry},\text{information}) = \frac{442*5+8*3982+2*3325}{\sqrt{442^2+8^2+2^2}\sqrt{5^2+3982^2+3325^2}} = .017$$

$$\cos(\text{digital},\text{information}) = \frac{5*5+1683*3982+1670*3325}{\sqrt{5^2+1683^2+1670^2}\sqrt{5^2+3982^2+3325^2}} = .996$$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Συχνότητα = Κακή εκπροσώπηση

- Οι πίνακες συν-εμφάνισης αναπαριστούν κάθε κύτταρο με συχνότητες λέξεων
- Η συχνότητα είναι χρήσιμη → Η ζάχαρη εμφανίζεται πολύ κοντά στο βερίκοκο.
- Υπερβολικά συχνές λέξεις όπως *the*, *it*, or *they* δεν είναι ενημερωτικά σχετικά με το πλαίσιο.
- Paradox: Πώς να εξισορροπήσετε αυτούς τους αντικρουόμενους περιορισμούς?

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Αναπαραστάσεις Ενσωμάτωσης Λέξεων

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Αναπαραστάσεις ενσωμάτωσης λέξεων

- Count-based
  - TF-IDF, PPMI
- Class-based
  - Brown Clusters
- Κατανεμημένη Πρόβλεψη με βάση Embeddings
  - Word2Vec, FastText
- Contextual Embeddings απο Γλωσσικά Μοντέλα
  - Elmo, BERT
- Παραλλαγές
  - Multi-lingual, Multi-sense etc.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

TF-IDF

- TF-IDF: Score λέξης *t* σε έγγραφο *d* $w_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t$
  - Λέξεις όπως "*the*" ή "*it*" έχουν χαμηλό IDF.

- PMI: Pointwise Mutual Information $\mathbf{PMI}(w_1, w_2) = log \frac{p(w_1,w_2)}{p(w_1)p(w_2)}$
  - Δείτε αν λέξεις όπως "good" εμφανίζονται συχνότερα με "great" από ό, τι θα περιμέναμε τυχαία

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

TF-IDF

- Term Frequency (TF):
  - $tf_{t,d} = count(t,d) \rightarrow tf_{t,d} = \log_{10}(count(t,d) + 1)$
- Document Frequency (DF):
  - df_t είναι ο αριθμός των εγγράφων στα οποία εμφανίζεται το t
- Inverse Document Frequency (IDF):
  - N είναι ο συνολικός αριθμός των εγγράφων της συλλογής

$$\text{idf}_t = \log_{10}\left(\frac{N}{\text{df}_t}\right)$$

Word	DF	IDF
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0.0
sweet	37	0.0

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

TF-IDF Word Weighted Values

- Raw counts:

$$w_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t$$

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

- TF-IDF:

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.049	0.044	0.018	0.022

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Sparse vs. Dense Vectors

- TF-IDF διανύσματα είναι:
  - **Long** length |V| = 20,000 to 50,000
  - **Sparse** most elements are zero
- Τα διανύσματα πρέπει να είναι:
  - **Short** length |V| = 200 to 1000
  - **Dense** most elements are non-zero
- **Short** vectors είναι ευκολότερα στη χρήση ως features
- **Dense** vectors είναι καλύτερα γενικευμένα
- Λειτουργούν καλύτερα στην πράξη.

Μέθοδοι για Short Dense Vectors

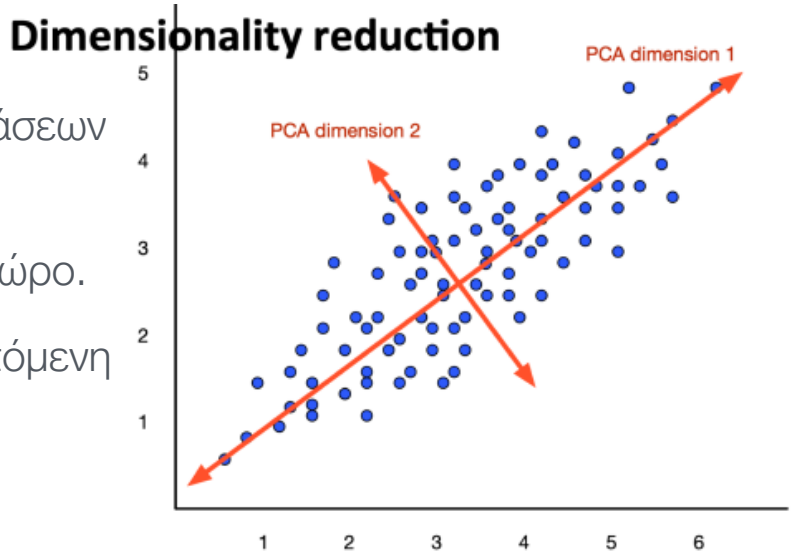
- Singular Value Decomposition (SVD)
  - LSA - Latent Semantic Analysis
- Brown Clustering
- Neural Language Model - Word2Vec

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Dense Vectors μέσω SVD

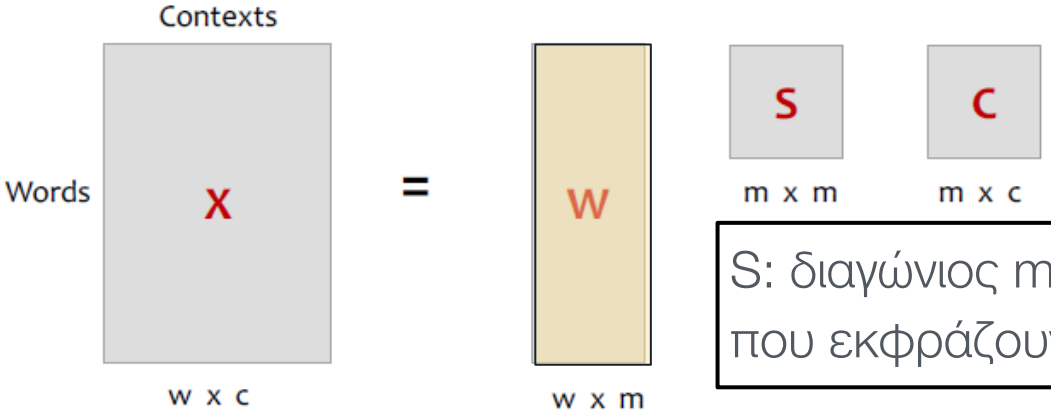
❑ Διαίσθηση

- Κατά προσέγγιση ένα σύνολο δεδομένων N-διαστάσεων χρησιμοποιώντας λιγότερες διαστάσεις.
- Περιστρέφοντας πρώτα τους άξονες σε ένα νέο χώρο.
- Η διάσταση υψηλότερης τάξης καταγράφει την επόμενη μεγαλύτερη διακύμανση κ.λπ.
- Υπάρχουν πολλές τέτοιες μέθοδοι:
  - PCA - Principal Components Analysis
  - Factor Analysis
  - SVD



Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Singular Value Decomposition SVD



S: διαγώνιος m x m πίνακας μοναδικών τιμών που εκφράζουν τη σημασία κάθε διάστασης

W: οι γραμμές που αντιστοιχούν στο αρχικό, αλλά οι στήλες m αντιπροσωπεύουν μια διάσταση σε ένα νέο λανθάνον χώρο, έτσι ώστε 1) , τα διανύσματα στηλών να είναι ορθογώνια μεταξύ τους και 2) οι στήλες να ταξινομούνται με βάση το μέγεθος της διακύμανσης στο σύνολο δεδομένων που αντιπροσωπεύει κάθε νέα διάσταση.

C: στήλες που αντιστοιχούν στο πρωτότυπο κατά m σειρές που αντιστοιχούν σε μοναδικές τιμές.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

SVD προς Term-Document Matrix
Latent Semantic Analysis

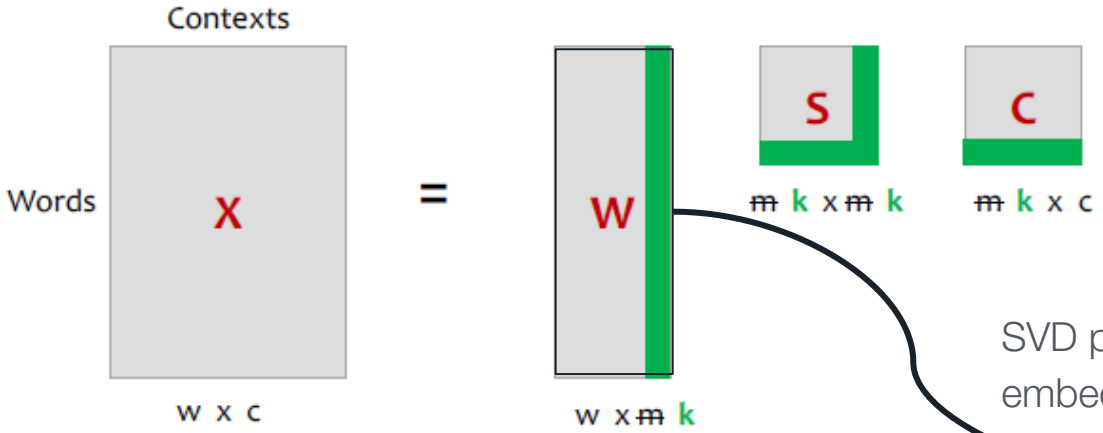
- Εάν αντί να διατηρήσουμε όλες τις διαστάσεις m, κρατάμε μόνο τα top-k singular values → Π.χ. 300

- Το αποτέλεσμα είναι μια προσέγγιση ελαχίστων τετραγώνων στο αρχικό X.

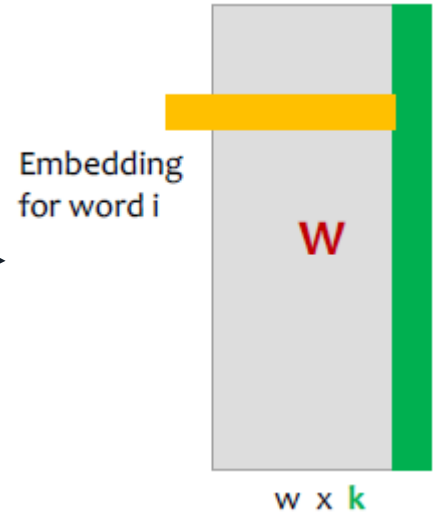
- Αντί όμως να πολλαπλασιάζεται → αξιοποιήστε το W.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Truncated SVD



SVD produces embeddings



- Κάθε σειρά W είναι ένα k-dimensional αναπαράσταση κάθε λέξης w.
- k μπορεί να κυμαίνεται από 50 προς 100
- Γενικά κρατάμε το top-k dimensions, αλλά τα πειράματα προτείνουν να απαλλαγούμε από το top-1 to 50 dimensions.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Embeddings vs. Sparse Vectors

- Οι πυκνές ενσωματώσεις SVD μερικές φορές λειτουργούν καλύτερα από τις αραιές μήτρες PPMI και TF-IDF σε εργασίες όπως η ομοιότητα λέξεων.

- Denoising: χαμηλές διαστάσεις → ασήμαντες πληροφορίες.
- Truncation μπορεί να βοηθήσει generalization to unseen data.
- Μικρότερος αριθμός διαστάσεων → ευκολότερη για classifiers Για να σταθμίσετε σωστά τις διαστάσεις της εργασίας.
- Dense models → καλύτερα στην καταγραφή συν-εμφανίσεων υψηλότερης τάξης.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Brown Clustering Algorithm

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

The Brown Clustering Algorithm

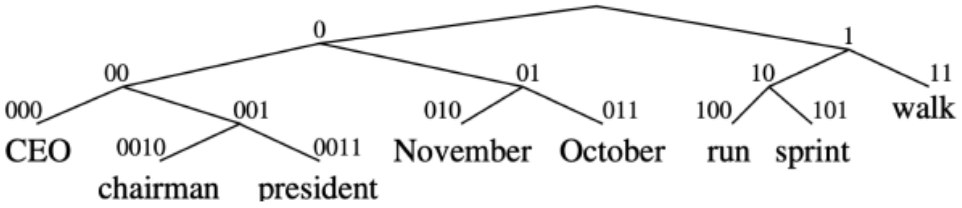
- **Input**: μια μεγάλη συλλογή λέξεων.
- **Output 1**: ένα partition των λέξεων σε word clusters.
- **Output 2**: ιεραρχική word clustering.

- Agglomerative clustering αλγόριθμος που ομαδοποιεί λέξεις με βάση τις λέξεις που προηγούνται ή τις ακολουθούν.
- Αυτά τα συμπλέγματα λέξεων μπορούν να μετατραπούν σε διανύσματα.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

The Brown Clustering Algorithm

- Κάθε λέξη αποδίδεται αρχικά στη δική της cluster.
- Τώρα εξετάζουμε τη συγχώνευση κάθε ζεύγους clusters.
  - Quality = Συγχωνεύει δύο λέξεις που έχουν παρόμοιες πιθανότητες να προηγηθούν και να ακολουθήσουν λέξεις.
- Clustering προχωρά μέχρι όλες οι λέξεις να είναι σε ένα μεγάλο cluster.
- Με την ανίχνευση της σειράς με την οποία clusters συγχωνεύονται, το μοντέλο δημιουργεί ένα δυαδικό δέντρο από κάτω προς τα πάνω.



"chairman" = 0010
"months" = 01
"verbs" = 1

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Simple Brown Algorithm

1. Ξεκινήστε με | *V* | clusters: Κάθε λέξη 1 cluster.
2. Ο στόχος είναι να πάρει *k* clusters.
3. Τρέχουμε | V | - k merge steps:

- Διαλέγω 2 clusters και να τα συγχωνεύσετε.
- Κάθε βήμα επιλέγει τη συγχώνευση μεγιστοποιώντας την ποιότητα του C

Cost: $O(|V| - k) \times O(|V|^2) \times O(|V|^2) = O(|V|^5)$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Word2Vec

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

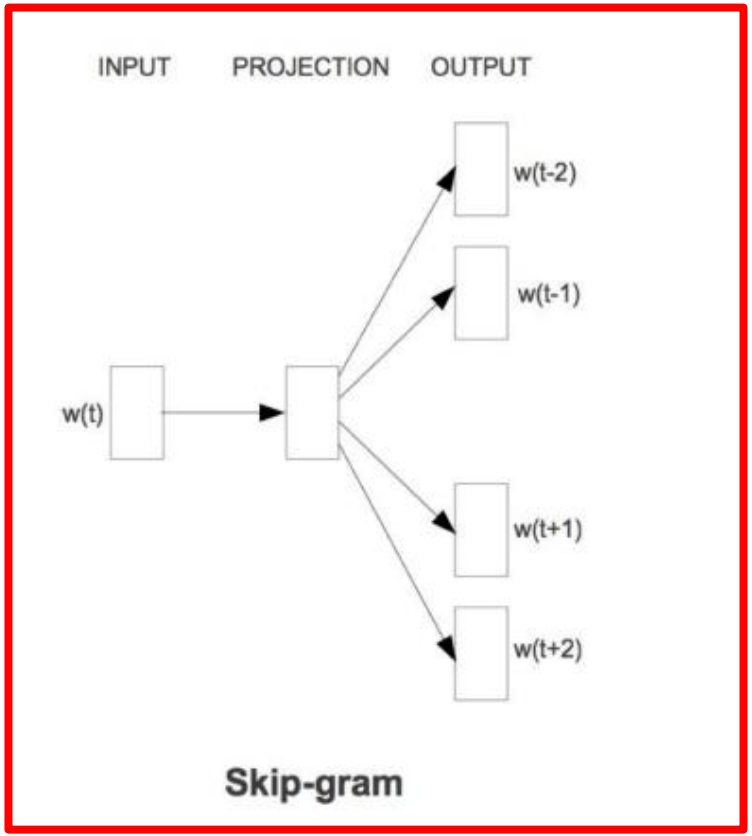
This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

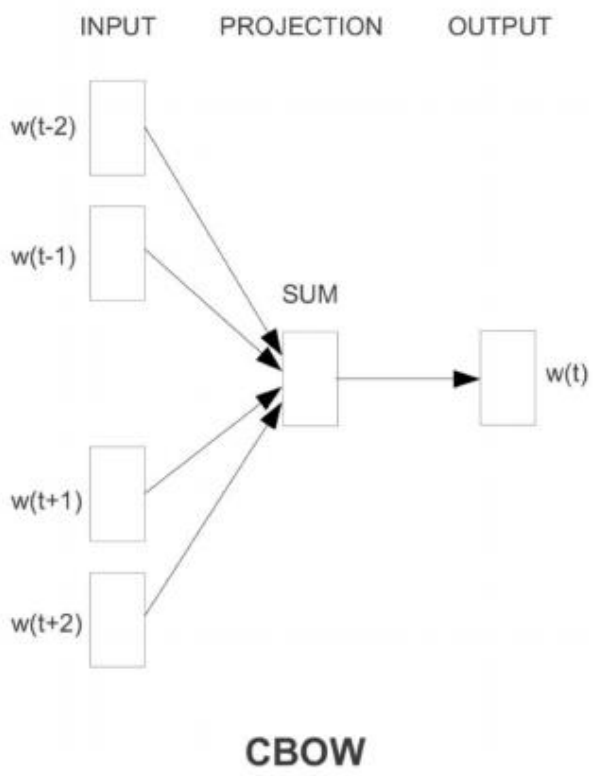
Word2Vec

- Δημοφιλής μέθοδος ενσωμάτωσης.
- Πολύ γρήγορο στην εκπαίδευση / υπολογισμό.
- **Ιδέα**: Πρόβλεψη αντί να μετράτε.
  - Παρόμοια με το γλωσσικό μοντέλο, αλλά η πρόβλεψη της επόμενης λέξης **ΔΕΝ** είναι ο στόχος.
- **Διαίσθηση**: Λέξεις που είναι σημασιολογικά παρόμοιες συχνά εμφανίζονται κοντά η μία στην άλλη στο κείμενο.
  - Οι ενσωματώσεις που είναι καλές στην πρόβλεψη γειτονικών λέξεων είναι επίσης καλές στην αναπαράσταση της ομοιότητας.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Word2Vec Πακέτο Αλγορίθμων





Θα εξετάσουμε

Skip-gram with Negative Sampling SGNS

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Εκπαίδευση με Αυτοεπίβλεψη

- Μια λέξη w που εμφανίζεται κοντά στη λέξη ροδάκινο στο σώμα λειτουργεί ως η χρυσή «σωστή απάντηση» για την εποπτευόμενη μάθηση.
  - *Είναι πιθανό η λέξη w να εμφανιστεί κοντά στο ροδάκινο?*
- Δεν υπάρχει ανάγκη επίβλεψης με ανθρώπινο σήμα.
- Η ιδέα προέρχεται από Neural Language Modeling
  - Bengio et al. 2003
  - Collobert et al. 2011

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Skip-Gram Sketch:
Προβλέψτε εάν ο υποψήφιος w είναι "γείτονας"

1. Αντιμετωπίστε τη λέξη-στόχο t και μια γειτονική λέξη περιβάλλοντος c ως **θετικά παραδείγματα**.
2. Τυχαία δειγματοληψία άλλων λέξεων στο λεξικό για λήψη αρνητικών δειγμάτων.
3. Χρησιμοποιήστε λογιστική παλινδρόμηση για να εκπαιδεύσετε έναν ταξινομητή να διακρίνει αυτές τις δύο περιπτώσεις.
4. Χρησιμοποιήστε τα μαθημένα βάρη ως ενσωματώσεις.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Skip-Gram Classification Task

- Θεωρείστε παράθυρο λέξεων *a* ± 2 και πρόταση:

  … lemon, a [tablespoon of apricot jam, a] pinch …

  c_1 c_2 [target] c_3 c_4

- Στόχος: Εκπαίδευση classifier με input ζευγάρι (**w**ord, **c**ontext):
  - (apricot, jam) P(+ | “*apricot*”, “*jam*”) → High
  - (apricot, aardvark) P(- | “*apricot*”, “*aardvark*”) → High

  …

  - Επιστρέφει την πιθανότητα κάθε ζεύγους:
    - P(+ | *w*, *c*)
    - P(- | *w*, *c*) = 1 - P(+ | *w*, *c*)

Base **P** on embedding similarity

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Embedding Similarity σε Probability

- Η ομοιότητα των δύο ενσωματώσεων w και c μπορεί να υπολογιστεί ως dot product
  - Similarity(w, c) ∝ w ▪ c

w ▪ c ∈ [-∞, +∞] Not a Probability

- Normalize $w \cdot c$: Μετατρέψτε την ομοιότητα σε πιθανότητα χρησιμοποιώντας Logistic Regression sigmoid σ.

$$P(+|w,c) = \sigma(c \cdot w) = \frac{1}{1+\exp(-c \cdot w)}$$

$$P(-|w,c) = 1 - P(+|w,c)$$

$$= \sigma(-c \cdot w) = \frac{1}{1+\exp(c \cdot w)}$$

Probability for 1 context word.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Skip-Gram για Multiple Context Words

- P(+ | w, c) υπολογίζει την πιθανότητα για ένα context word.

$$P(+|w,c) = \sigma(c \cdot w) = \frac{1}{1+\exp(-c \cdot w)}$$

- Για πολλαπλές context words Υποθέτουμε ανεξαρτησία και πολλαπλασιάζουμε αυτές τις πιθανότητες.

- Για λέξεις L στο context window, πολλαπλασιάστε τα σιγμοειδή τους.

$$P(+|w,c_{1:L}) = \prod_{i=1}^{L} \sigma(c_i \cdot w)$$

$$\log P(+|w,c_{1:L}) = \sum_{i=1}^{L} \log \sigma(c_i \cdot w)$$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Skip-Gram Classifier: Περίληψη

- Μια πιθανοτική classifier, δεδομένο
  - Μια δοκιμαστική λέξη-στόχος w
    Context window των λέξεων L $c_{1:L}$
- Υπολογίζει την πιθανότητα εμφάνισης του w σε αυτό το παράθυρο με βάση την ομοιότητα των embeddings w σε $c_{1:L}$ embeddings.
- Για να το υπολογίσουμε αυτό χρειαζόμαστε ενσωματώσεις για τη λέξη-στόχο και τις λέξεις περιβάλλοντος.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Μαθαίνοντας τα Embeddings

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Εκπαίδευση του Skip-Gram Classifier

- Λάβετε υπόψη τα ακόλουθα δεδομένα εκπαίδευσης Skip-Gram:

  … lemon, a [tablespoon of apricot jam, a] pinch …

  c_1 c_2 [target] c_3 c_4

- Για κάθε θετικό παράδειγμα, παίρνουμε αρνητικά παραδείγματα με βάση τη συχνότητα.

Positive Examples +	
t	***c***
apricot	tablespoon
apricot	of
apricot	jam
apricot	a

Negative Examples -			
t	***c***	***t***	***c***
apricot	aardvark	apricot	seven
apricot	my	apricot	forever
apricot	where	apricot	dear
apricot	coaxial	apricot	if

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Μαθαίνοντας Word2Vec Vectors

- Input: θετικά και αρνητικά παραδείγματα κατάρτισης
- Output: Διανύσματα λέξεων 300 διαστάσεων
- Initialization: Τυχαία προετοιμασία των διανυσμάτων λέξεων.

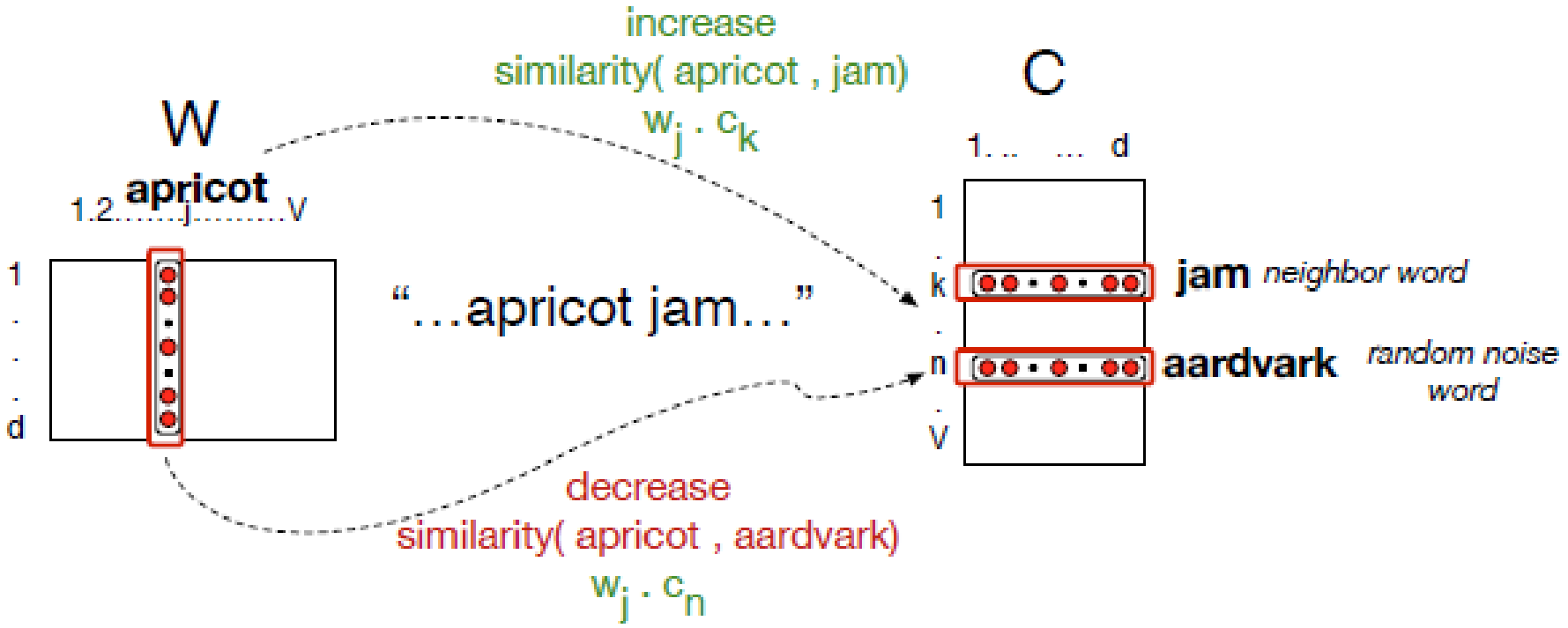
Εφαρμόστε επαναληπτικά τον μαθησιακό στόχο, προσαρμόζοντας τα διανύσματα έτσι ώστε να:

- **Μεγιστοποιήστε την ομοιότητα των** target και context ζευγών λέξεων (w, c_{POS}) από τα θετικά δεδομένα.
- **Ελαχιστοποιήστε την ομοιότητα των** (w, c_{NEG}) ζευγών λέξεων από τα αρνητικά δεδομένα

$$\sum_{(t,c)\in+} logP(+|t,c) + \sum_{(t,c)\in-} logP(-|t,c)$$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Διαίσθηση του Gradient Descent



Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Loss Function για *w*

Μεγιστοποιήστε την ομοιότητα του στόχου με τις πραγματικές λέξεις περιβάλλοντος και ελαχιστοποιήστε την ομοιότητα του στόχου με τις k αρνητικές μη γειτονικές λέξεις.

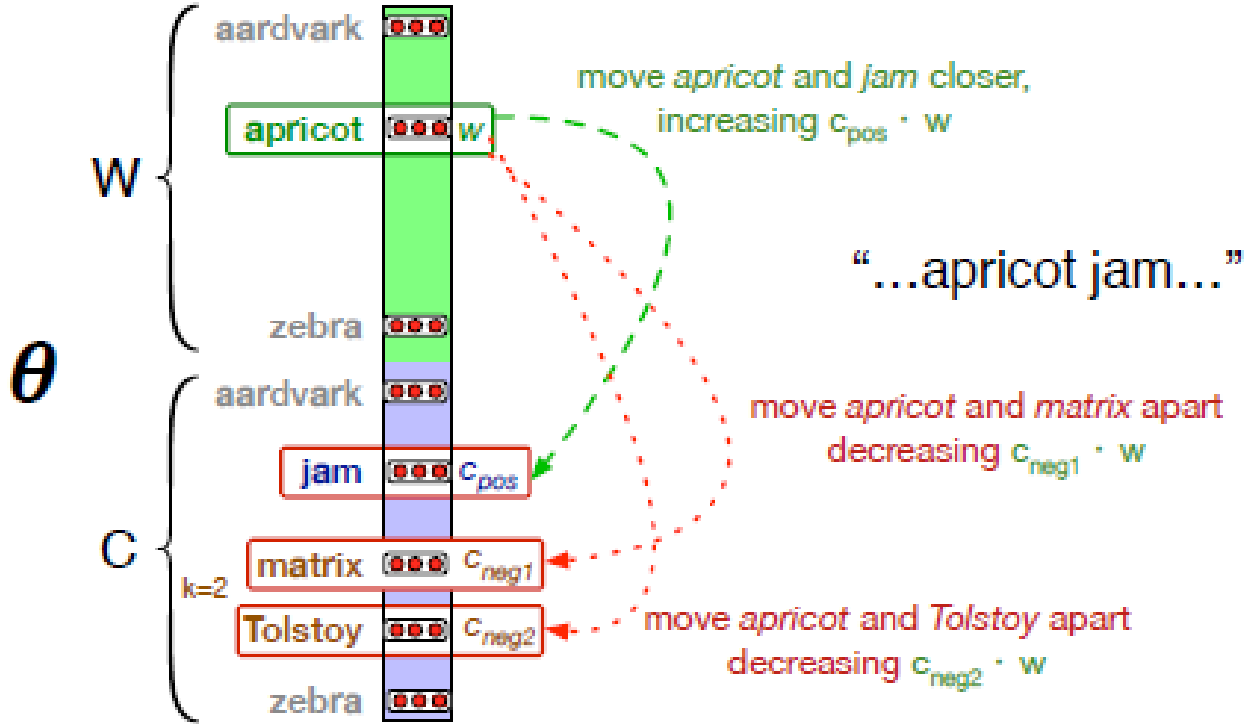
Ελαχιστοποιώ L_{CE} με Gradient Descent.
Προσαρμόστε το word weights έτσι ωστε

- Κάντε τα θετικά ζεύγη πιο πιθανά.
  Κάντε τα αρνητικά ζεύγη λιγότερο πιθανά.
  Σε ολόκληρο το σετ προπόνησης.

$$
\begin{aligned}
L_{CE} &= -\log\left[P(+|w,c_{pos})\prod_{i=1}^{k}P(-|w,c_{neg_i})\right] \\
&= -\left[\log P(+|w,c_{pos}) + \sum_{i=1}^{k}\log P(-|w,c_{neg_i})\right] \\
&= -\left[\log P(+|w,c_{pos}) + \sum_{i=1}^{k}\log\left(1-P(+|w,c_{neg_i})\right)\right] \\
&= -\left[\log \sigma(c_{pos}\cdot w) + \sum_{i=1}^{k}\log\sigma(-c_{neg_i}\cdot w)\right]
\end{aligned}
$$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Intuition του Gradient Descent



Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

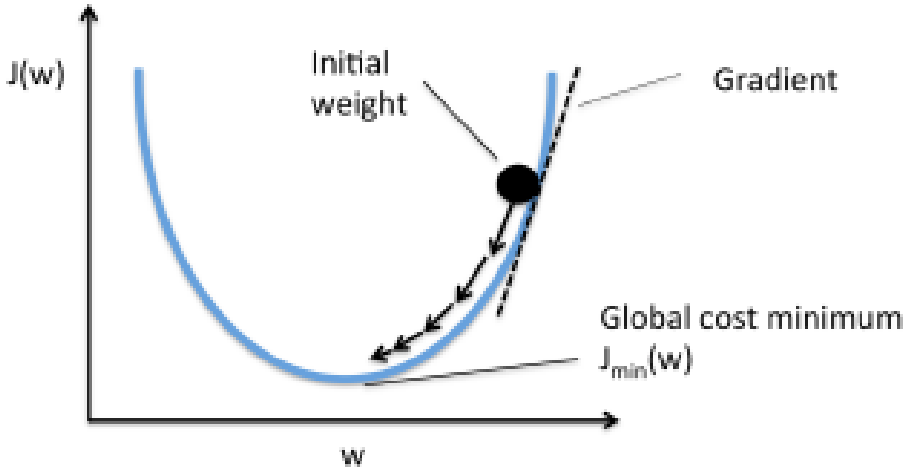
Gradient Descent

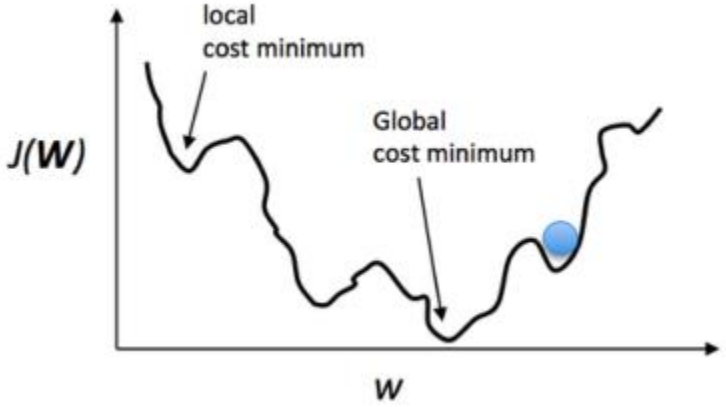
- Σε κάθε βήμα εκπαίδευσης:
- Κατεύθυνση: Κινούμαστε προς την αντίστροφη κατεύθυνση από την κλίση της συνάρτησης απώλειας.
- Μέγεθος: Μετακινούμε την αξία του gradient $\frac{d}{dw} L(f(x;w),y)$ σταθμισμένο από ένα learning rate η.
  - Υψηλότερος ρυθμός μάθησης σημαίνει ταχύτερη κίνηση w.

$$w^{t+1} = w^t - \eta \frac{d}{dw} L(f(x;w),y)$$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Gradient Descent





Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Derivatives of the Loss Function

Τα παράγωγα της συνάρτησης απώλειας:

$$L_{CE} = -\left[\log \sigma(c_{pos} \cdot w) + \sum_{i=1}^{k} \log \sigma(-c_{neg_i} \cdot w)\right]$$

Παράγωγο του Οι ενσωματώσεις περιβάλλοντος

$$\begin{cases} \dfrac{\partial L_{CE}}{\partial c_{pos}} = [\sigma(c_{pos} \cdot w) - 1]w \\ \dfrac{\partial L_{CE}}{\partial c_{neg}} = [\sigma(c_{neg} \cdot w)]w \end{cases}$$

Παράγωγος των ενσωματώσεων στόχου

$$\frac{\partial L_{CE}}{\partial w} = [\sigma(c_{pos} \cdot w) - 1]c_{pos} + \sum_{i=1}^{k} [\sigma(c_{neg_i} \cdot w)$$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

SGD Update Equation

- Ξεκινήστε με τυχαία αρχικοποιημένους πίνακες C και W και, στη συνέχεια, κάντε σταδιακά ενημερώσεις.

$$c_{pos}^{t+1} = c_{pos}^{t} - \eta[\sigma(c_{pos}^{t} \cdot w^{t}) - 1]w^{t}$$

$$c_{neg}^{t+1} = c_{neg}^{t} - \eta[\sigma(c_{neg}^{t} \cdot w^{t})]w^{t}$$

$$w^{t+1} = w^{t} - \eta\left[[\sigma(c_{pos} \cdot w^{t}) - 1]c_{pos} + \sum_{i=1}^{k}[\sigma(c_{neg_i} \cdot w^{t})]c_{neg_i}\right]$$

- Skipgram μαθαίνει δύο σετ embeddings: 1) target embeddings matrix *W*; και 2) context embedding matrix *C*.
  - Προσθέστε τα μαζί, αντιπροσωπεύοντας τη λέξη i ως διάνυσμα $w_i + c_i$

Μάθηση Word2Vec: Περίληψη

- Ξεκινήστε με V τυχαία d-διαστάσεων διανύσματα ως αρχικές ενσωματώσεις.
  Εκπαίδευση ταξινομητή με βάση την ομοιότητα ενσωμάτωσης;
  - Πάρτε ένα σώμα κειμένων και πάρτε ζεύγη λέξεων που συνυπάρχουν ως θετικά παραδείγματα.
    Πάρτε ζεύγη λέξεων που δεν συνυπάρχουν ως αρνητικά παραδείγματα.
    Εκπαιδεύστε τον ταξινομητή να τα διακρίνει προσαρμόζοντας αργά όλες τις ενσωματώσεις για να βελτιώσετε την απόδοση του ταξινομητή.
    Πετάξτε τον κωδικό ταξινομητή και συνεχίστε τις ενσωματώσεις;

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Embeddings Properties και Evaluation

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

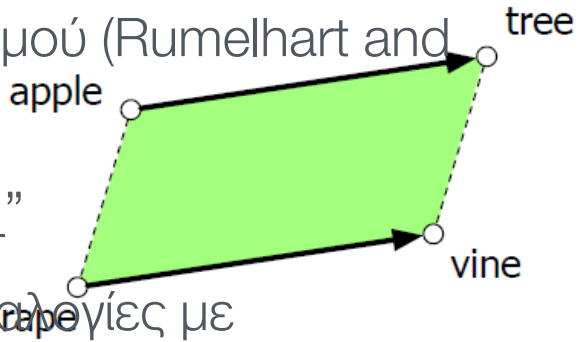
Γείτονες και μέγεθος παραθύρου

- Μικρά παράθυρα ($C = \pm 2$): Οι πλησιέστερες λέξεις είναι συντακτικά παρόμοιες στην ίδια ταξινόμηση.
  - Οι πλησιέστεροι γείτονες του Hogwards είναι άλλα φανταστικά σχολεία, π.χ. Sunnydale, Evernight, Blandings κ.λπ.
- Μεγάλα παράθυρα ($C = \pm 5$): Οι πλησιέστερες λέξεις είναι σχετικές λέξεις στο ίδιο σημασιολογικό πεδίο.
  - Οι πλησιέστεροι γείτονες του Χόγκουαρντ σχετίζονται με τον Χάρι Πότερ, δηλαδή τον Ντάμπλντορ, τον ημίαιμο, τον Μαλφόι κ.λπ.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Analogy: Σχεσιακή έννοια

- Μοντέλο παραλληλόγραμμου αναλογικού συλλογισμού (Rumelhart and Abrahamson 1973).
- *Πρόβλημα*: "***Apple*** *is to* ***tree*** *as* ***grape*** *is to* _____"
- Η μέθοδος παραλληλόγραμμου μπορεί να λύσει αναλογίες με τόσο αραιές όσο και πυκνές ενσωματώσεις (Mikolov et al. 2013)



vector("*king*") - vector("*man*") + vector("*woman*") ≈ vector("*queen*")
vector("Paris") - vector("*France*") + vector("*Italy*") ≈ vector("*Rome*")

- For a problem a:a*::b:b*, the parallelogram method is:

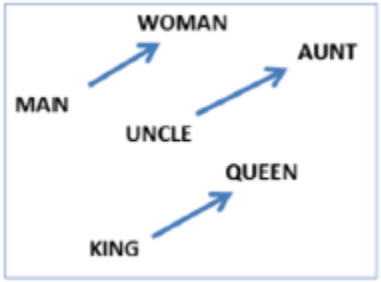
$$\hat{b}^* = \underset{x}{\text{argmax}}\ \text{distance}(x, a^* - a + b)$$

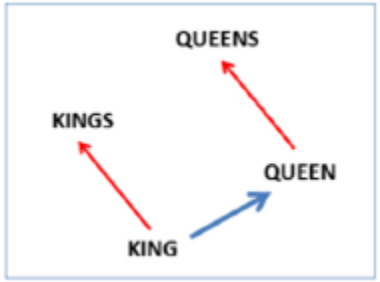
Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

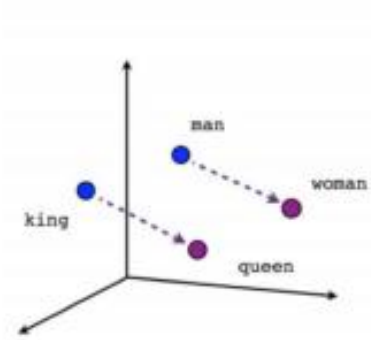
Analogy: Relational Meaning

vector("*king*") - vector("*man*") + vector("*woman*") ≈ vector("*queen*")

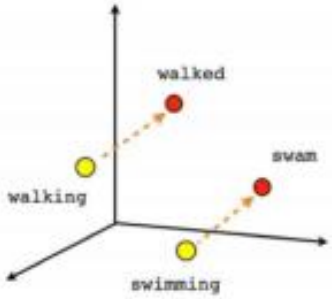
vector("Paris") - vector("*France*") + vector("*Italy*") ≈ vector("*Rome*")



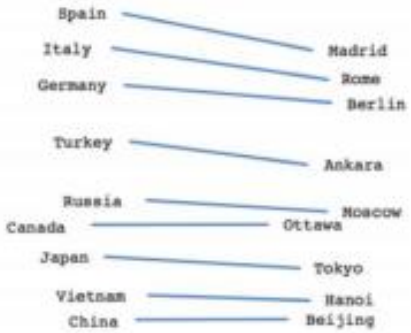




Male-Female



Verb tense



Country-Capital

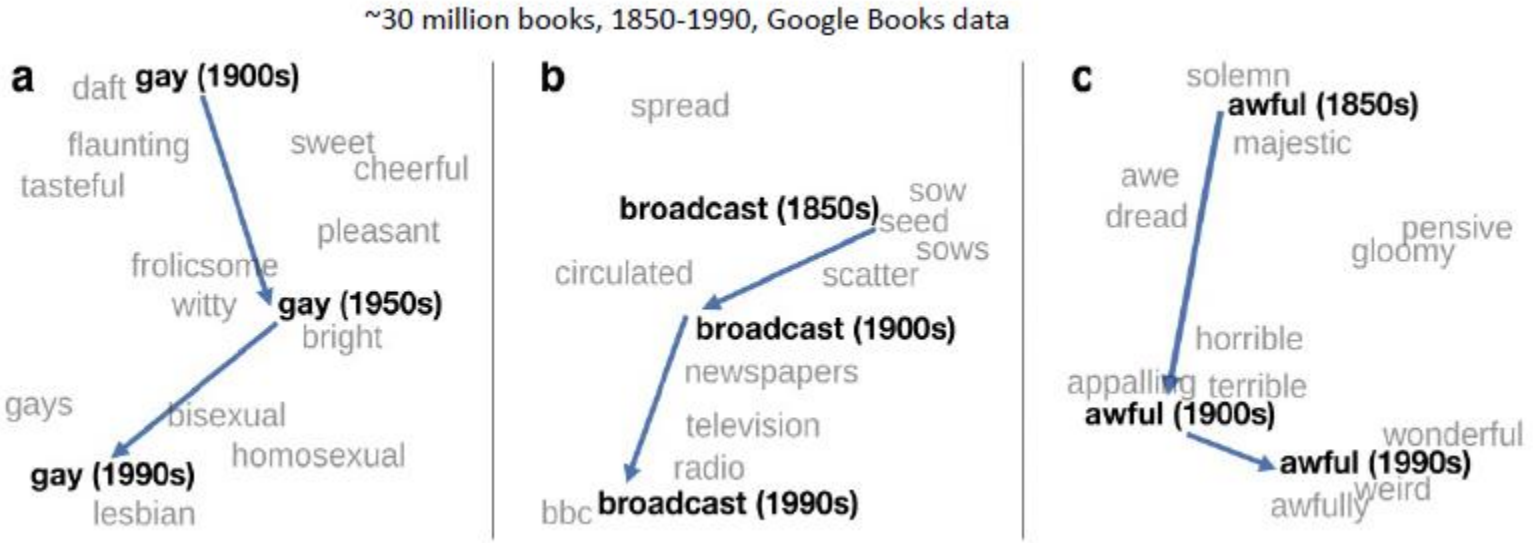
Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Embeddings & Historical Semantics

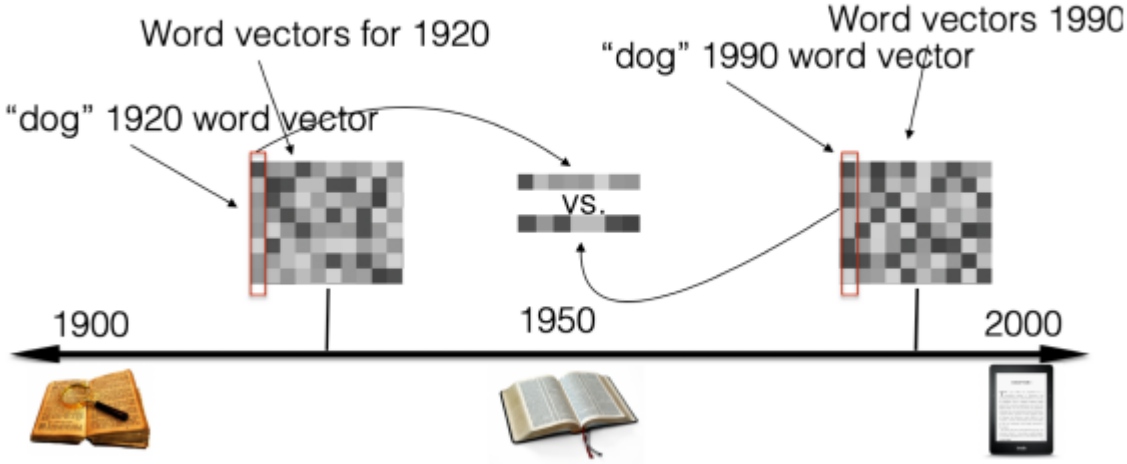
- Εκπαιδεύστε ενσωματώσεις σε διαφορετικές δεκαετίες ιστορικού κειμένου για να δείτε τα νοήματα να αλλάζουν (Hamilton et al. 2016) .

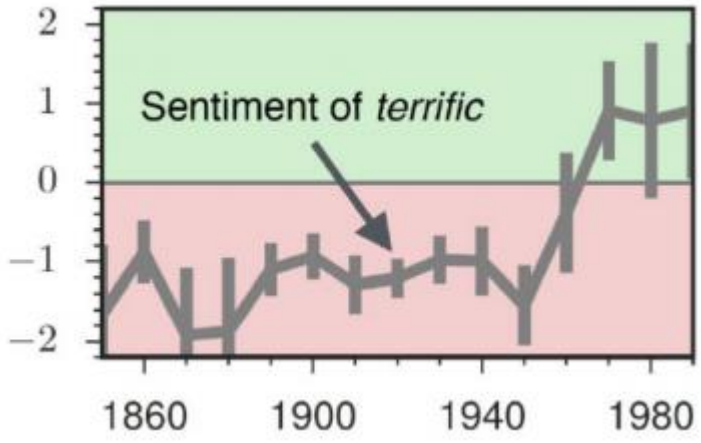


Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Embeddings & Historical Semantics

- Εκπαιδεύστε ενσωματώσεις σε διαφορετικές δεκαετίες ιστορικού κειμένου για να δείτε τα νοήματα να αλλάζουν (Hamilton et al. 2016).





Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Embeddings & Πολιτισμική προκατάληψη

- Οι ενσωματώσεις αντικατοπτρίζουν πολιτισμική προκατάληψη (Bolukbasi et al. 2016).
- Ask “*Paris* : *France* :: *Tokyo* : X”
  - X = “*Japan*”
- Ask “*father* : *doctor* :: *mother* : X”
  - X = “*nurse*”
- Ask “*man* : *computer programmer* :: *woman* : X”
  - X = “*homemaker*”

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Embeddings & Πολιτισμική προκατάληψη

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

***softball* extreme**	**gender portion**	**after debiasing**
1. pitcher	-1%	1. pitcher
2. bookkeeper	20%	2. infielder
3. receptionist	67%	3. major leaguer
4. registered nurse	29%	4. bookkeeper
5. waitress	35%	5. investigator

***football* extreme**	**gender portion**	**after debiasing**
1. footballer	2%	1. footballer
2. businessman	31%	2. cleric
3. pundit	10%	3. vice chancellor
4. maestro	42%	4. lecturer
5. cleric	2%	5. midfielder

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Embeddings & Πολιτισμική προκατάληψη

- Δοκιμή έμμεσης συσχέτισης (Greenwald et al 1998): How associated are:
  - Έννοιες (λουλούδια, έντομα) &; ιδιότητες (ευχαρίστηση, δυσάρεστη);
    Μελετήθηκε μετρώντας λανθάνοντες χρόνους για κατηγοριοποίηση.
- Ψυχολογικά ευρήματα για τους συμμετέχοντες στις ΗΠΑ:
  - Τα αφροαμερικανικά ονόματα συνδέονται με δυσάρεστες λέξεις (περισσότερο από τις ευρωπαϊκές-αμερικανικές)
    Τα αρσενικά ονόματα συνδέονται περισσότερο με τα μαθηματικά, τα γυναικεία ονόματα με τις τέχνες.
    Ονόματα ηλικιωμένων με δυσάρεστες λέξεις, νέοι με ευχάριστες λέξεις.
- Caliskan et al. 2017 αναπαραγωγή με embeddings:
  - Τα αφροαμερικανικά ονόματα (Leroy) είχαν υψηλότερο συνημίτονο GloVe με κακοποίηση, βρωμιά, άσχημο.
    Τα ευρωπαϊκά αμερικανικά ονόματα (Brad, Greg) είχαν υψηλότερο συνημίτονο με αγάπη, ειρήνη, θαύμα.
- Embeddings αντανακλούν και αναπαράγουν κάθε είδους ολέθριες προκαταλήψεις.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Περίληψη

- Παρουσίαση της Διανεμητικής Υπόθεσης, θεμελιώδους σημασίας των Ενσωματώσεων Λέξεων
- Υπολογιστικές Σημασιολογικές Προσεγγίσεις:
  - Count-based
  - TF-IDF
  - Class-based e.g. Brown Clusters
  - Word2Vec
  - Contextual Embeddings e.g BERT
- Word Semantic Similarity χρησιμοποιώντας Cosine Similarity

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Πόροι

- Jurafsky, D. and H. Martin Justin, Chapter 6. "Vector Semantics and Embeddings" Speech and Language Processing
- Python SKLearn TF-IDF Vectorization: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
- Python Gensim W2V: https://radimrehurek.com/gensim/models/word2vec.html
- Python Tutorial on Using Pre-trained W2V models: https://colab.research.google.com/github/practical-nlp/practical-nlp/blob/master/Ch3/05_Pre_Trained_Word_Embeddings.ipynb?authuser=1

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423