Εμφυτεύματα Λέξεων (Word Embeddings)

Βαθιά Μηχανική Μάθηση

ΔΠΜΣ Επιστήμης Δεδομένων & Μηχανικής Μάθησης

Εθνικό Μετσόβιο Πολυτεχνείο

Γιώργος Αλεξανδρίδης (gealexan@mail.ntua.gr)

Αναπαράσταση Κειμένων

- *Πως* μπορεί να αναπαρασταθεί *αριθμητικά* ένα κείμενο;
  - Βασικό *αιτούμενο* αλλά και *στάδιο* του επιστημονικού πεδίου της **Επεξεργασίας Φυσικής Γλώσσας** (*Natural Language Processing – NLP*)
- Μια *ορθή* και *αποδοτική* αριθμητική αναπαράσταση μας επιτρέπει να αυτοματοποιήσουμε πλήθος σχετικών εργασιών
  - *μετάφραση*, *δημιουργία περιλήψεων, ταξινόμηση εγγράφων, συσταδοποίηση εγγράφων*, …
- Ένα κείμενο μπορεί να αποτελείται από *κεφάλαια*, *ενότητες*, *παραγράφους*, *λέξεις*, *γράμματα* και *σημεία στίξης*

Ελάχιστη Ελεύθερη Μορφή

- **Leonard Bloomfield** (1887-1949)
  - Αμερικανός Γλωσσολόγος, συνέβαλλε στην ανάπτυξη του πεδίου της δομικής γλωσσολογίας τις δεκαετίες του 1930-1940 στις ΗΠΑ
  - Το 1928 ανέπτυξε την έννοια της **Ελάχιστης Ελεύθερης Μορφής** (*Minimal Free Form*), σύμφωνα με την οποία οι *λέξεις* είναι η *ελάχιστη ενότητα λόγου*, η οποία μπορεί *αυτοτελώς* να βγάλει νόημα

- Πως μπορούμε να αναπαραστήσουμε αριθμητικά τις λέξεις;
  1. Ως **διακριτά σύμβολα** (*discrete symbols*)
  2. Μέσω του **συγκειμένου** (*context*)

Αναπαράσταση λέξεων ως διακριτά σύμβολα

- **Κωδικοποίηση one-hot** (*one-hot encoding*)
  - Κάθε λέξη και *διαφορετική* διάσταση
  - Παράδειγμα

    Ξενοδοχείο = [0 0 0 0 1 0 0 0 0 0]

    Πανδοχείο = [1 0 0 0 0 0 0 0 0 0]

    Σκύλος = [0 1 0 0 0 0 0 0 0 0]

- *Μήκος* διανύσματος είναι ίσο προς το *μέγεθος* του *λεξιλογίου*
  - πχ 1.000.000 λέξεις

- Πως μπορεί να υπολογιστεί η *σημασιολογική ομοιότητα* μεταξύ των λέξεων;

Αναπαράσταση λέξεων από το συγκείμενο

- **John Rupert Firth** (1890-1960)
  - Κορυφαίος Βρετανός Γλωσσολόγος
  - "*You shall know a word by the company it keeps*"
    - 1957, *Studies in Linguistic Analysis*, Wiley-Blackwell
  - Από τις πιο επιτυχημένες προσεγγίσεις στη σύγχρονη στατιστική ΕΦΓ!
- **Υπόθεση Κατανομής**
  - *Λέξεις* που προκύπτουν σε *παρόμοια συγκείμενα* τείνουν να έχουν *παρόμοιο* νόημα
  - Στο παρακάτω παράδειγμα, λέξεις *συγκειμένου* που καθορίζουν την αναπαράσταση της λέξης ***banking***

…government debt problems turning into	**banking**	*crises as happened in 2009…*
…saying that Europe needs unified	**banking**	*regulation to replace the hodgepodge…*
…India has just given its	**banking**	*system a shot in the arm…*

Υπόθεση Κατανομής

«*Λέξεις που προκύπτουν σε παρόμοια συγκείμενα τείνουν να έχουν παρόμοιο νόημα*»

C_1: Ένα μπουκάλι _____ βρίσκεται στο τραπέζι

C_2 : Σε πολλούς αρέσει το _____

C_3 : Μην πίνετε _____ πριν οδηγήσετε

C_4: Το _____ μπορεί να περιέχει γλυκάνισο

	C_1	C_2	C_3	C_4
τσίπουρο	1	1	1	1
λάδι	1	1	0	0
μηχανή	0	0	0	0
ψωμί	0	1	0	0
πετρέλαιο	1	0	0	0
κρασί	1	1	1	0

Οι λέξεις ως διανύσματα

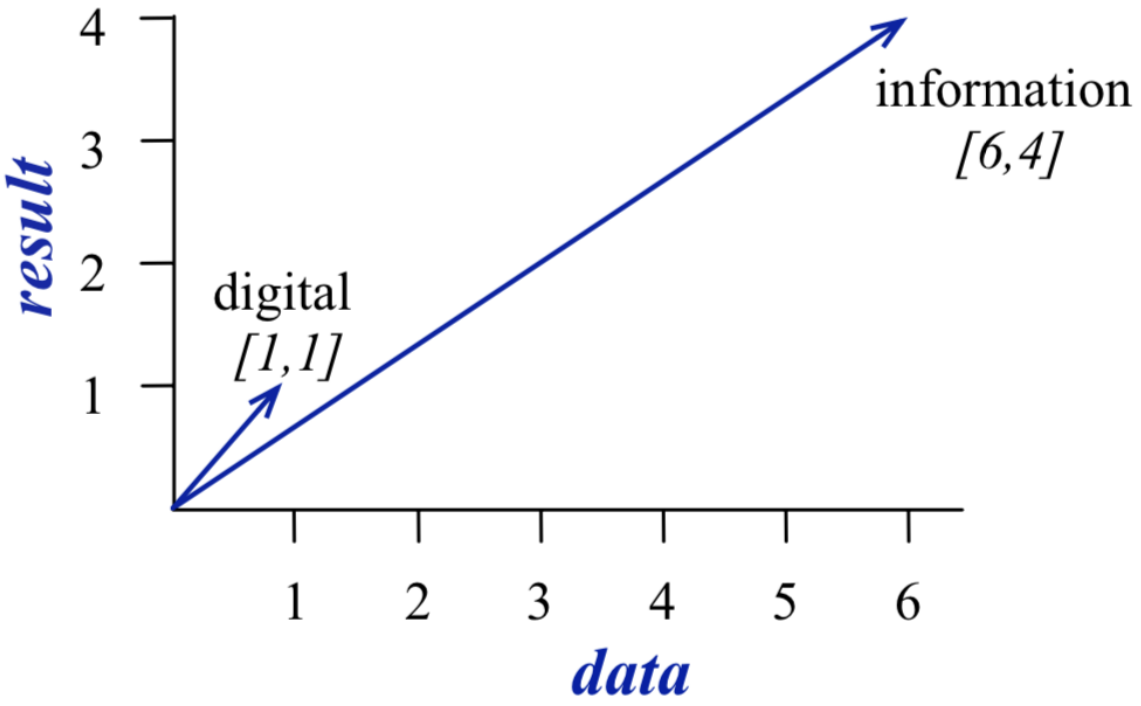
- Επιθυμούμε να κατασκευάσουμε ένα *μοντέλο* βασισμένο στην *ομοιότητα*
  - Κάθε λέξη αναπαρίσταται ως *διάνυσμα*
  - *Παρόμοιες λέξεις* απεικονίζονται σε *κοντινές θέσεις* στο χώρο

- Μια πρώτη προσέγγιση
  - Χρήση διανυσμάτων *συγκειμένου* για την αναπαράσταση του νοήματος των λέξεων

Πίνακας συνεμφάνισης λέξεων

	aardvark	computer	data	pinch	result	sugar	…
apricot	0	0	0	1	0	1	
pineapple	0	0	0	1	0	1	
digital	0	2	1	0	1	0	
information	0	1	6	0	4	0	

Οι λέξεις ως διανύσματα



Ομοιότητα συνημίτονου

$$\cos(\boldsymbol{u}, \boldsymbol{v}) = \frac{\boldsymbol{u}.\boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|}$$

Πρόβλημα το διαφορετικό μήκος των διανυσμάτων

Διαφορετικό μήκος διανυσμάτων

	computer	data	result	pie	sugar
cherry	2	8	9	442	25
strawberry	0	0	1	60	19
digital	1670	1683	85	5	4
information	3325	3982	378	5	13

- **<u>Λύση</u>**: *Στάθμιση* της συχνότητας εμφάνισης λέξης με τεχνικές όπως η **σημειακή αμοιβαία πληροφορία** (*positive pointwise mutual information - PPMI*)

$$PPMI(word, context) = \max\left(\log_2 \frac{P(word, context)}{P(word)P(context)}, 0\right)$$

	computer	data	result	pie	sugar
cherry	0	0	0	4.38	3.30
strawberry	0	0	0	4.10	5.51
digital	0.18	0.01	0	0	0
information	0.02	0.09	0.28	0	0

Πυκνότητα Διανυσμάτων

- Τα διανύσματα που παίρνουμε είναι **αραιά** (*sparse*)
  - *Μεγάλα* σε μέγεθος (ίσο με το μέγεθος του λεξικού)
  - Πλειοψηφία των στοιχείων τους είναι *μηδενικά*
- Ιδανικά θα θέλαμε διανύσματα *μικρών διαστάσεων* (πχ $50 - 300$) και **πυκνά** (*dense*)
  - τιμές τους να είναι *μη-μηδενικοί* πραγματικοί αριθμοί όπως στο διπλανό παράδειγμα

$$\text{employees} = \begin{pmatrix} 0{,}286 \\ 0{,}792 \\ -0{,}177 \\ -0{,}107 \\ 10{,}109 \\ -0{,}542 \\ 0{,}349 \\ 0{,}271 \\ 0{,}487 \end{pmatrix}$$

Word2vec

- *Μέθοδος* εκμάθησης *αναπαράστασης* λέξεων ως *πυκνά* διανύσματα
  - Mikolov et al (2013), *Distributed Representations of Words and Phrases and their Compositionality*

- <u>Είσοδος</u>
  1. Μια μεγάλη συλλογή κειμένου
     - πχ Wikipedia + Gigaword, Twitter, Common Crawl, …
  2. Λεξικό V
  3. Εύρος d διανύσματος λέξεων (π.χ. 300)

- <u>Έξοδος</u>
  - $f: V \rightarrow \mathbb{R}^d$

$$v_{cat} = \begin{pmatrix} -0{,}224 \\ 0{,}130 \\ -0{,}290 \\ 0{,}276 \end{pmatrix} \quad v_{dog} = \begin{pmatrix} -0{,}124 \\ 0{,}430 \\ -0{,}200 \\ 0{,}329 \end{pmatrix}$$

$$v_{the} = \begin{pmatrix} 0{,}234 \\ 0{,}266 \\ 0{,}239 \\ -0{,}199 \end{pmatrix} \quad v_{language} = \begin{pmatrix} 0{,}290 \\ -0{,}441 \\ 0{,}762 \\ 0{,}982 \end{pmatrix}$$

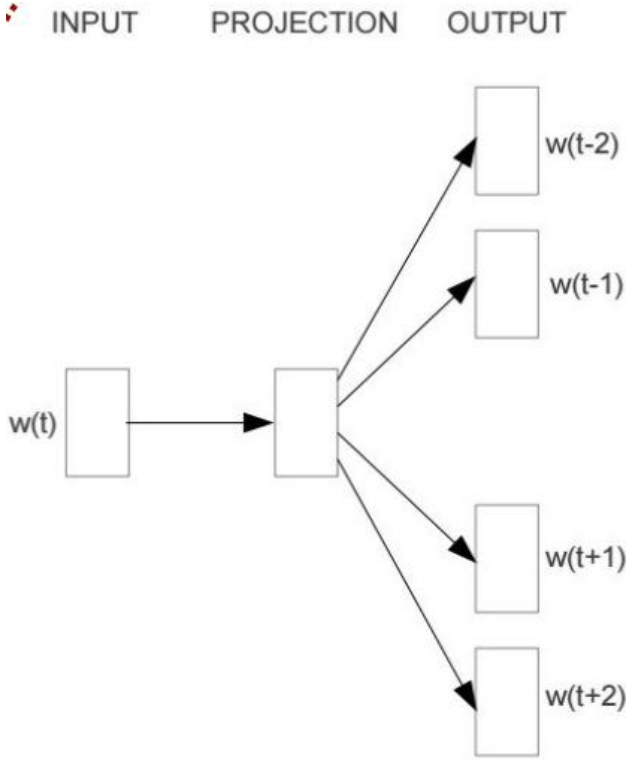
Word2vec: ομοιότητα με λέξη “sweden”

Word	Cosine distance
norway	0.760124
denmark	0.715460
finland	0.620022
switzerland	0.588132
belgium	0.585835
netherlands	0.574631
iceland	0.562368
estonia	0.547621
slovenia	0.531408

Word2vec: Τρόποι λειτουργίας

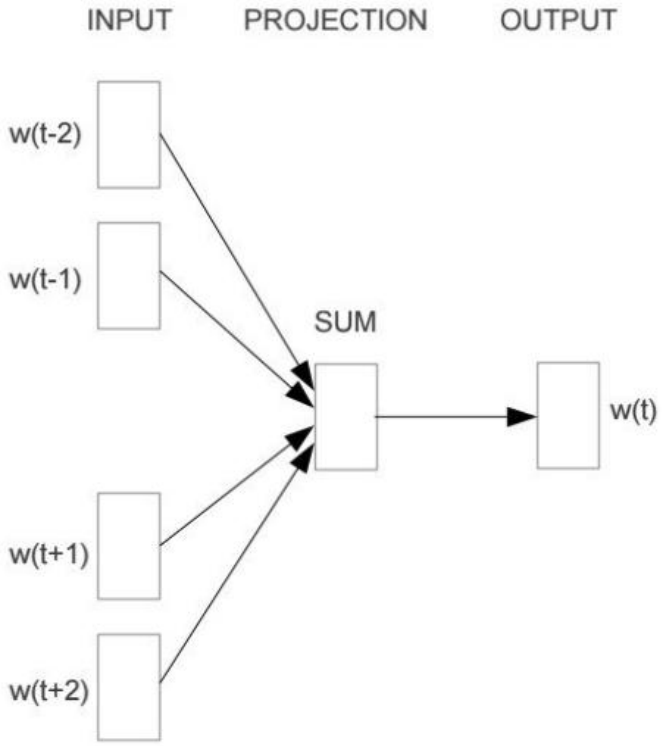
Skip-gram

- Πρόβλεψε τις *συγκείμενες* λέξεις μιας δεδομένης *λέξης-στόχου*



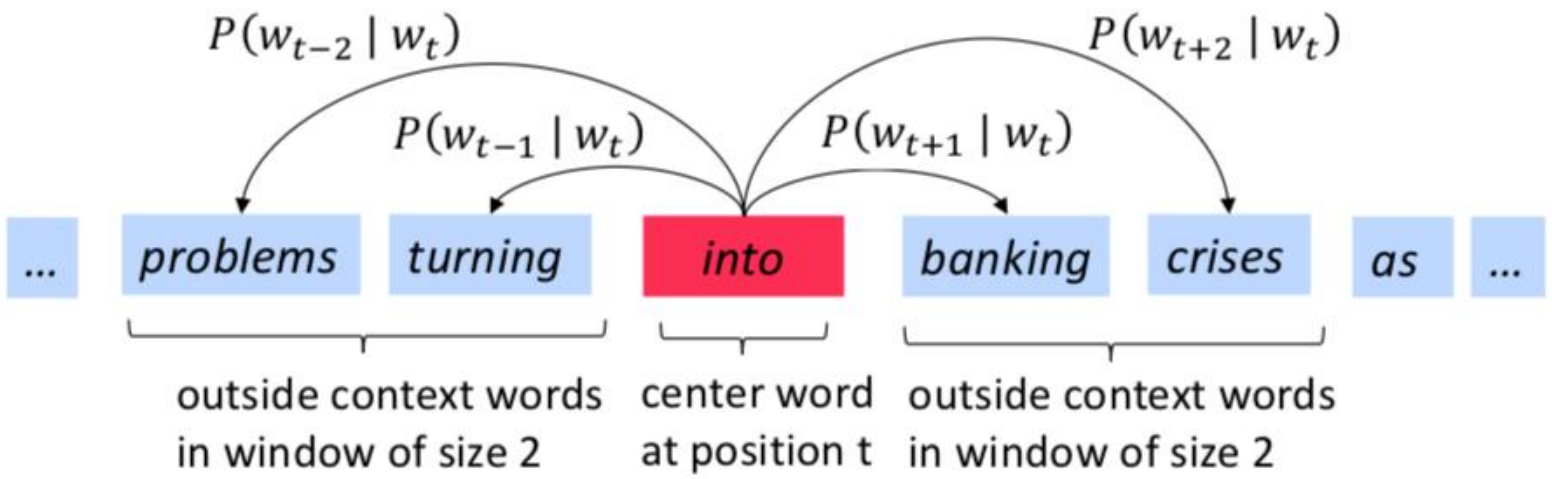
Continuous Bag-of-Word (CBOW)

- Πρόβλεψε τη λέξη που *ταιριάζει* στο συγκείμενο



Λειτουργία Skip-gram

- Θέλουμε να χρησιμοποιήσουμε τις λέξεις για να προβλέψουμε τις συγκείμενες τους (σταθερό μέγεθος παραθύρου $2m$)



Αντικειμενική Συνάρτηση

- Για κάθε θέση $t = 1,2, \dots, T$ στο κείμενο υπολόγισε την πιθανοφάνεια συγκειμένων λέξεων πλήθους m γύρω από την κεντρική λέξη w_j

$$\mathcal{L}(\theta) = \prod_{t=1}^{T} \prod_{-m \le j \le m, j \neq 0} P(w_{t+j} | w_t; \theta)$$

- Ως αντικειμενική συνάρτηση παίρνουμε τη μέση αρνητική πιθανοφάνεια

$$J(\theta) = -\frac{1}{\mathrm{T}} \log \mathcal{L}(\theta) = -\frac{1}{\mathrm{T}} \sum_{t=1}^{T} \sum_{-m \le j \le m, j \neq 0} \log P(w_{t+j} | w_t; \theta)$$

Ορισμός υπό-συνθήκη πιθανοτήτων

- Δύο σύνολα διανυσμάτων ή διαφορετικά σύνολα **εμφυτευμάτων** (*embeddings*)
  - $\boldsymbol{u}_i \in \mathbb{R}^d$ εμφύτευμα για τη λέξη i όταν είναι λέξη-στόχος
  - $\boldsymbol{v}_{i'} \in \mathbb{R}^d$ εμφύτευμα για τη λέξη i' όταν είναι συγκείμενη λέξη
  - Χρήση *εσωτερικού γινομένου* $\boldsymbol{u}_i.\boldsymbol{v}_{i'}$ για τον προσδιορισμό του πόσο πιθανό είναι να εμφανίζεται η λέξη i ως συγκείμενη της λέξης i'
- $P\left(w_{t+j}\middle|w_t;\theta\right) = \frac{e^{\boldsymbol{u}_{w_t}.\boldsymbol{v}_{w_{t+j}}}}{\sum_{k\in V} e^{\boldsymbol{u}_{w_t}.\boldsymbol{v}_k}} = \mathrm{s}\left(\boldsymbol{u}_{w_t}.\boldsymbol{v}_{w_{t+j}}\right)$
  - συνάρτηση softmax $\mathrm{s}(x_i) = \frac{e^{x_i}}{\sum_{i=1}^{n} e^{x_i}}$
- $\theta = \{\{\boldsymbol{u}_k\},\{\boldsymbol{v}_k\}\}$, οι εκπαιδεύσιμες παράμετροι του μοντέλου

Εκπαίδευση μοντέλου

- **Στοχαστική Κατάβαση Κλίσης** (*Stochastic Gradient Descend – SGD*)
  - $\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta \nabla_\theta J(\theta)$
- **Αντικειμενική Συνάρτηση**
  - $J(\theta) = -\frac{1}{\mathrm{T}} \sum_{t=1}^{T} \sum_{-m \le j \le m, j \neq 0} \log P\left(w_{t+j} \middle| w_t; \theta\right) = -\frac{1}{\mathrm{T}} \sum_{t=1}^{T} \sum_{-m \le j \le m, j \neq 0} \log \frac{e^{\boldsymbol{u}_{w_t} \cdot \boldsymbol{v}_{w_{t+j}}}}{\sum_{k \in V} e^{\boldsymbol{u}_{w_t} \cdot \boldsymbol{v}_k}}$
  - $J(\{\{\boldsymbol{u}_k\}, \{\boldsymbol{v}_k\}\}) = -\frac{1}{\mathrm{T}} \sum_{t=1}^{T} \sum_{-m \le j \le m, j \neq 0} [\boldsymbol{u}_{w_t} . \boldsymbol{v}_{w_{t+j}} - \log(\sum_{k \in V} e^{\boldsymbol{u}_{w_t} \cdot \boldsymbol{v}_k})]$
- Υπολογίζουμε την *κλίση* για ένα *συγκεκριμένο* ζεύγος *λέξης-στόχου* t και *λέξης συγκειμένου* c
  - $\frac{\partial J}{\partial \boldsymbol{u}_t} = \frac{\partial(-\boldsymbol{u}_t . \boldsymbol{v}_c + \log(\sum_{k \in V} e^{\boldsymbol{u}_t \cdot \boldsymbol{v}_c}))}{\partial \boldsymbol{u}_t} = -\boldsymbol{v}_c + \frac{\sum_{k \in V} e^{\boldsymbol{u}_t \cdot \boldsymbol{v}_k} \boldsymbol{v}_k}{\sum_{k \in V} e^{\boldsymbol{u}_t \cdot \boldsymbol{v}_k}} = -\boldsymbol{v}_c + \sum_{k \in V} P(k|t)\, \boldsymbol{v}_k$
  - Αντίστοιχα $\frac{\partial J}{\partial \boldsymbol{u}_t} = -1_{k=c} \boldsymbol{u}_t + P(k|t) \boldsymbol{u}_t$

Εκπαίδευση μοντέλου (συνέχεια)

- Είσοδος
  - Συλλογή κειμένων, παράθυρο συγκειμένου m, διάσταση διανύσματος εμφυτευμάτων d, λεξικό V
- Αλγόριθμος Εκπαίδευσης
  1. Αρχικοποίηση $u_i, v_i, i \in V$ σε τυχαίες τιμές
  2. Επεξεργασία των κειμένων της συλλογής σειριακά για τον εντοπισμό ζευγών λέξεων στόχου t και συγκειμένου c
     a. $\boldsymbol{u}_t{}' \leftarrow \boldsymbol{u}_t - \eta \frac{\partial J}{\partial \boldsymbol{u}_t} \Rightarrow \boldsymbol{u}_t{}' \leftarrow \boldsymbol{u}_t + \eta \boldsymbol{v}_c - \eta \sum_{k \in V} P(k|t)\, \boldsymbol{v}_k$
     b. $\boldsymbol{v}_k{}' \leftarrow \boldsymbol{v}_k - \eta \frac{\partial J}{\partial \boldsymbol{v}_k} \Rightarrow \boldsymbol{v}_k{}' \leftarrow \boldsymbol{v}_k + \eta_{k=c}\, \boldsymbol{u}_t - \eta P(k|t) \boldsymbol{u}_t, \forall k \in V$
- Ζήτημα Υλοποίησης
  - Για την επεξεργασία ενός μόνο ζεύγους λέξεων στόχου-συγκειμένου πρέπει να ενημερωθούν όλα τα εμφυτεύματα συγκειμένου στο λεξικό!

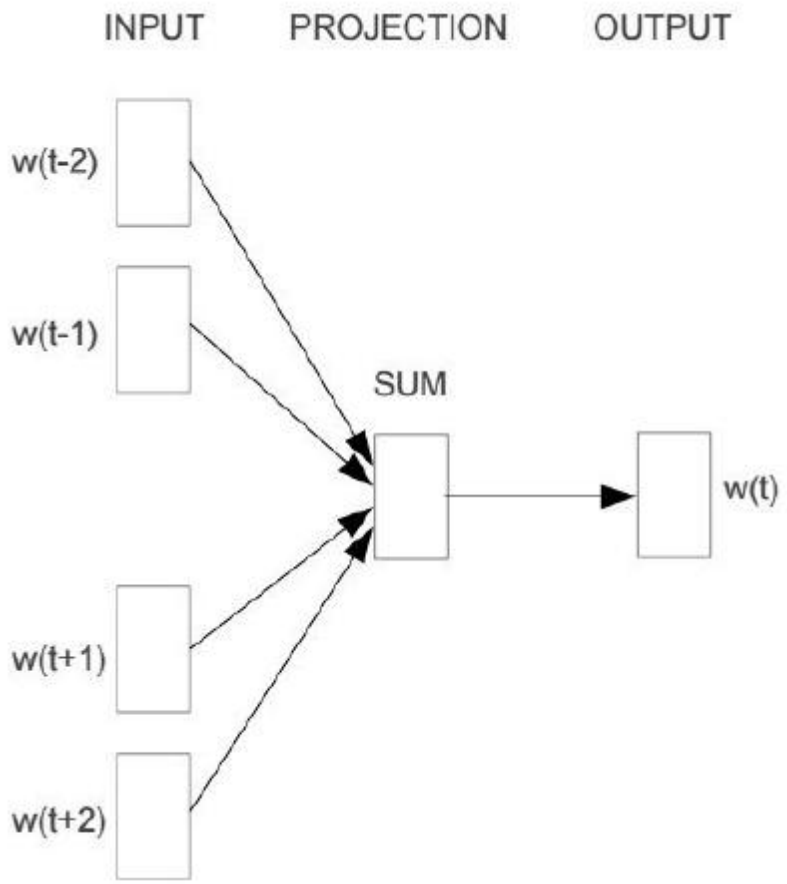
Λύση: Αρνητική Δειγματοληψία!

- *Skip-gram with Negative Sampling* (*SGNS*)
- Αντί να ενημερώσουμε όλα τα εμφυτεύματα συγκειμένου, *δειγματοληπτούμε K* (5-20) **αρνητικά παραδείγματα** (*negative samples*)
  - Αρνητικά παραδείγματα: Λέξεις που δεν είναι συγκείμενες της εκάστοτε λέξης-στόχου
- Τροποποίηση αντικειμενικής συνάρτησης
  - $J(\boldsymbol{u}_t, \boldsymbol{v}_c) = -\log(\sigma(\boldsymbol{u}_t.\boldsymbol{v}_c)) - \sum_{i=1}^{K} \mathbb{E}_{j\sim P(w)} \log(-\sigma(\boldsymbol{u}_t.\boldsymbol{v}_c))$
    - σιγμοειδής συνάρτηση $\sigma(x) = \frac{1}{1+e^{-x}}$
  - Ομοιότητα με εκπαίδευση λογιστικής παλινδρόμησης ($P(D = 1|t, c) = \sigma(\boldsymbol{u}_t.\boldsymbol{v}_c)$)

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

Λειτουργία Continuous Bag of Words (CBOW)



- <u>Αντικειμενική συνάρτηση</u>
  - $L(\theta) = \prod_{t=1}^{T} P(w_t | \{w_{t+j}\}_{-m \le j \le m, j \neq 0})$
- Προσδιορισμός *υπό συνθήκη πιθανότητας* με χρήση συνάρτησης *softmax*
  - $P\left(w_t \middle| \{w_{t+j}\}_{-m \le j \le m, j \neq 0}\right) = s(\boldsymbol{u}_{w_t} \cdot \overline{\boldsymbol{v}_t})$
- $\overline{\boldsymbol{v}_t}$: Μέση τιμή *εμφυτευμάτων* συγκειμένου
  - $\overline{\boldsymbol{v}_t} = \frac{1}{2m} \sum_{-m \le j \le m, j \neq 0} \boldsymbol{v}_{t+j}$

GloVe: Global Vectors

- *Μέθοδος* εκμάθησης *αναπαράστασης* λέξεων ως *πυκνά* διανύσματα
  - Pennington et al (2014), *GloVe: Global Vectors for Word Representation*
- Βασίζεται στον υπολογισμό της *συχνότητας συνεμφάνισης* των λέξεων στα κείμενα εκπαίδευσης
- <u>Αντικειμενική συνάρτηση</u>
  - $J = \sum_{i,j=1}^{V} f(X_{ij})\ (w_i^T \widetilde{w_j} + b_i + \widetilde{b_j} - \log X_{ij})^2$
  - X_{ij}, συχνότητα συνεμφάνισης λέξεων i, j
- <u>Πλεονεκτήματα</u> (σε σύγκριση με Word2vec)
  - *Ταχύτερη* εκπαίδευση
  - *Κλιμακωσιμότητα* σε πολύ μεγάλες συλλογές δεδομένων

GloVe: Παράδειγμα

Nearest words to frog:

1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

litoria

leptodactylidae

rana

eleutherodactylus

FastText

- Υπολογισμός εμφυτευμάτων στο επίπεδο της «**υπο-λέξης**» (*sub-word*)
  - Bojanowski et al (2017): *Enriching Word Vectors with Subword Information*
- Όμοια λειτουργία με το skip-gram, με τη διαφορά να είναι το σπάσιμο των λέξεων σε n-grams, όπου $n \in \{3,4,5,6\}$
- Παράδειγμα λέξης **where**
  - 3-grams: <wh, wher, her, ere, re>
  - 4-grams: <whe, wher, here, ere>
  - 5-grams: <where,where>
- Αντικατάσταση εμφυτευμάτων $\boldsymbol{u}_i$ συγκειμένου από το άθροισμα των n-grams
  - Αντί για τον υπολογισμού του $\boldsymbol{u}_i \cdot \boldsymbol{v}_j$, υπολόγισε το $\sum_{g \in n-grams(w_i)} \boldsymbol{u}_g \cdot \boldsymbol{v}_j$

Βιβλιογραφία

- Word2Vec
  - Mikolov et al (2013), *Distributed Representations of Words and Phrases and their Compositionality*
- Glove
  - Pennington et al (2014), *GloVe: Global Vectors for Word Representation*
- Fast Text
  - Bojanowski et al (2017), *Enriching Word Vectors with Subword Information*
- Παρουσίαση Word Embeddings
  - COS 484: Natural Language Processing, Πανεπιστήμιο Princeton
  - Σε αυτήν βασίστηκε σε μεγάλο βαθμό η τρέχουσα παρουσίαση