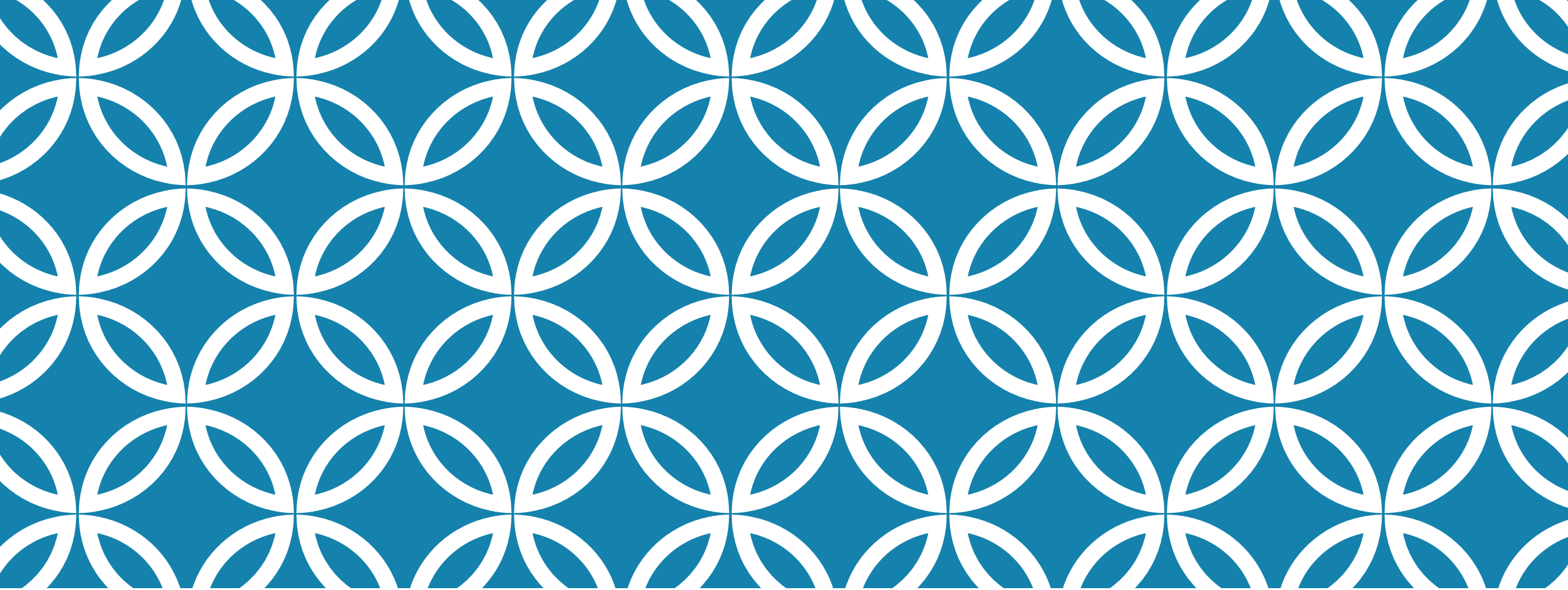

ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

Κάτια Κερμανίδου kerman@ionio.gr

LANGUAGE MODELING

Το Μοντέλο Γλώσσας (Language Model)

Το Μοντέλο Γλώσσας

- Προβλέπει ποια μπορεί να είναι η επόμενη λέξη σε μια ακολουθία λέξεων

Το νερό στη θάλασσα της Γλυφάδας είναι ιδιαιτέρως …

… καθαρό
… διαυγές
--- κρύο

…. τροχάδην
… δεδομένου
… καναπές

- Αποδίδει πιθανότητες σε ακολουθίες λέξεων (πόσο πιθανό είναι να εμφανιστούν στη γλώσσα)

Οι εκπρόσωποι αποφάσισαν κοινή πολιτική.

Οι εκπρόσωποι αποφάσισε κοινή πολιτική.

Το Μοντέλο Γλώσσας (Language Model)

Ένα καλό Μοντέλο Γλώσσας πρέπει

- να δίνει μεγάλη πιθανότητα σε προτάσεις που είναι συχνές/σωστές προτάσεις σε μια γλώσσα
- να δίνει χαμηλή πιθανότητα σε προτάσεις που είναι σπάνιες/λανθασμένες προτάσεις σε μια γλώσσα

Μοντέλο Γλώσσας: Χρησιμότητα

- Στην Μηχανική Μετάφραση

  Ποτέ δεν το είπα αυτό I never said that I never not that said that

- Στην Σύνθεση Φυσικής Γλώσσας

  LLMs, Διαλογικά συστήματα

- Στην Αναγνώριση Ομιλίας

  «…Ήρθα []ωρίς το σακίδιό μου…» → …Ήρθα χωρίς το σακίδιό μου…

- Στον γραμματικό/ορθογραφικό έλεγχο

  Η μάνα ψήνεις μπισκότα → Η μάνα ψήνει μπισκότα

Μοντέλο Γλώσσας: Πιο Μαθηματικά

- Υπολογισμός της πιθανότητας μιας ακολουθίας n λέξεων

  $P(w_1, w_2, w_3, \ldots, w_n)$

- Υπολογισμός της πιθανότητας εμφάνισης της επόμενης λέξης w_n, δεδομένης μιας ακολουθίας n-1 λέξεων

  $P(w_n \mid w_1, w_2, w_3, \ldots, w_{n-1})$

Πώς υπολογίζω αυτές τις πιθανότητες; Αν χρησιμοποιήσω ένα τεράστιο σώμα κειμένων και υπολογίσω

$P(w_n \mid w_1, w_2, w_3, \ldots, w_{n-1}) = \text{count}(w_1, w_2, w_3, \ldots, w_n) / \text{count}(w_1, w_2, w_3, \ldots, w_{n-1}) =$

count(“Το νερό στη θάλασσα της Γλυφάδας είναι ιδιαιτέρως καθαρό”) / count(“το νερό στη θάλασσα της Γλυφάδας είναι ιδιαιτέρως”)

Δεν θα βρω εμφανίσεις των ακολουθιών. Οι πιθανές ακολουθίες είναι πάρα πολλές!

Πώς υπολογίζονται οι πιθανότητες;

Chain Probabilities

$P(w_1, w_2, w_3, \ldots, w_n) = P(w_1) \times P(w_2|w_1) \times P(w_3|w_1, w_2) \times \ldots \times P(w_n| w_1, w_2, w_3, \ldots, w_{n-1})$

πχ

P("το νερό στη θάλασσα") = P(το) x P(νερό | το) x P(στη | το νερό) x P(θάλασσα | το νερό στη)

Όσο μεγαλώνει η ακολουθία των λέξεων, οι πιθανότητες όσο πηγαίνω προς τα δεξιά του τύπου γίνονται όλο και πιο σπάνιες, και από ένα σημείο και μετά μηδενικές.

Οπότε: Κάνω παραδοχές (assumptions) που απλοποιούν τους παραπάνω υπολογισμούς.

Παραδοχή N-γραμμου (N-gram Markov Assumption): Μια λέξη σε μια ακολουθία λέξεων δεν εξαρτάται από όλες τις προηγούμενες λέξεις της ακολουθίας, αλλά μόνο από τις προηγούμενες N-1 λέξεις.

Παραδοχή N-γράμμου

Αντί για

$P(w_1, w_2, w_3, \ldots, w_n) = P(w_1) \times P(w_2|w_1) \times P(w_3|w_1, w_2) \times \ldots \times P(w_n| w_1, w_2, w_3, \ldots, w_{n-1})$

Θεωρώ ότι

$P(w_1, w_2, w_3, \ldots, w_n) \approx P(w_1) \times P(w_2|w_1) \times P(w_3|w_1, w_2) \times \ldots \times P(w_n| w_{n-N+1}, w_{n-N+2}, w_{n-N+3}, \ldots, w_{n-1})$

N=4 (τετράγραμμο)

$P(w_1, w_2, w_3, \ldots, w_n) \approx P(w_1) \times P(w_2|w_1) \times P(w_3|w_1, w_2) \times \ldots \times P(w_n| w_{n-3}, w_{n-2}, w_{n-1})$

N=3 (τρίγραμμο)

$P(w_1, w_2, w_3, \ldots, w_n) \approx P(w_1) \times P(w_2|w_1) \times P(w_3|w_1, w_2) \times \ldots \times P(w_n| w_{n-2}, w_{n-1})$

N=2 (Bigram - δίγραμμο)

$P(w_1, w_2, w_3, \ldots, w_n) \approx P(w_1) \times P(w_2|w_1) \times P(w_3|w_2) \times \ldots \times P(w_n| w_{n-1})$

N=1 (Unigram – μονόγραμμο)

$P(w_1, w_2, w_3, \ldots, w_n) \approx P(w_1) \times P(w_2) \times P(w_3) \times \ldots \times P(w_n)$

Παραδείγματα

Παραδείγματα Διγράμμου

Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one gram point five percent of U. S. E. has already old M. X. corporation of living

I hire the men who is good pilots

Παραδείγματα Μονογράμμου

To him swallowed confess hear both . Which . Of save on trail for are ay device and rote life have

Hill he late speaks ; or ! a more to leg less first you enter

Πώς υπολογίζω τις πιθανότητες δεξιά στον τύπο;

Εκτιμήσεις Μέγιστης Πιθανοφάνειας (Maximum Likelihood Estimates)

$P(w_n|w_1, w_2, \ w_{n-1}) = count(w_1, w_2, \ldots, w_n) / count(w_1, w_2, \ldots, w_{n-1})$

N=4

$P(w_4|w_1, w_{2,} w_3) = count(w_1, w_2, w_{3,} w_4) / count(w_1, w_{2,} w_3)$

N=3

$P(w_3|w_1, w_2) = count(w_1, w_2, w_3) / count(w_1, w_2)$

N=2

$P(w_2|w_1) = count(w_1, w_2) / count(w_1)$

Παράδειγμα Υπολογισμών

<s> είμαι ο Γιάννης </s>

<s> ο Γιάννης είμαι </s>

<s> μου αρέσει ο τραχανάς </s>

P(ο) = 3/ 16

P(ο|<s>) = count("<s> ο") / count(<s>) = 1/3

P(Γιάννης | ο) = count("ο Γιάννης") / count(ο) = 2/3

Προβλήματα με τα N-γραμμα

Έλλειψη γνώσης σχετικά με την δομή της πρότασης.
Δεν μπορούν να αντιμετωπίσουν εξαρτήσεις μεγάλων αποστάσεων (long distance dependencies)

Οι συνομιλίες στην Βουλή, βάσει των οποίων προέκυψε το συγκεκριμένο αίτημα, υπήρξαν ο βασικός λόγος για αυτή την απόφαση του Προέδρου.

Η λύση: τα Μεγάλα Μοντέλα Γλώσσας (Large Language Models – LLMs)

Αξιολόγηση του μοντέλου γλώσσας

Εξωγενής αξιολόγηση (Extrinsic Evaluation)

Μέσα από κάποια τρίτη εφαρμογή, όπως η Μηχανική Μετάφραση ή η Αναγνώριση Ομιλίας, για να συγκρίνω δυο μοντέλα γλώσσας, LMA και LMB.

1. Χρησιμοποιώ το LMA και μεταφράζω κείμενο ή μεταγράφω ομιλία.
2. Χρησιμοποιώ το LMB και μεταφράζω το ίδιο κείμενο ή μεταγράφω την ίδια ομιλία.
3. Μετράω πόσες λέξεις μεταφράστηκαν/μεταγράφηκαν σωστά με το LMA, πόσες με το LMB
4. Συγκρίνω την απόδοση

Μειονεκτήματα της εξωγενούς αξιολόγησης

Εξαρτάται από την τρίτη εφαρμογή – δε μπορεί πάντα να επεκταθεί η αξιολόγηση σε άλλες

Ακριβή, χρονοβόρα

Αξιολόγηση του μοντέλου γλώσσας

Ενδογενής αξιολόγηση (Intrinsic Evaluation)
Αξιολογεί απευθείας το μοντέλο στην πρόβλεψη λέξεων (+)
Δεν αντιστοιχεί πάντα με την πραγματική απόδοση του μοντέλου σε εφαρμογές (-)
Δίνει ένα ενιαίο μαθηματικό μέτρο: Perplexity

Μεθοδολογία Αξιολόγησης
Σε σώμα εκπαίδευσης (training set) υπολογίζονται οι παράμετροι του μοντέλου (όλες οι πιθανότητες)
Σε σώμα ελέγχου/δοκιμής (test set) εφαρμόζονται οι υπολογισμένες πιθανότητες προκειμένου να υπολογισθεί το Perplexity.

Σώματα Εκπαίδευσης και Δοκιμής

Αν θέλω να χτίσω ένα Μοντέλο Γλώσσας για κάποια ειδική εργασία
και το σώμα εκπαίδευσης και το δοκιμής πρέπει να είναι αντιπροσωπευτικά της γλώσσας και της θεματικής περιοχής της εργασίας

Αν θέλω να χτίσω ένα Μοντέλο Γλώσσας γενικής χρήσης
και το σώμα εκπαίδευσης και το δοκιμής πρέπει να είναι διαφορετικών θεματικών περιοχών, διαφορετικών συγγραφέων και ποικίλου γλωσσολογικού ύφους

Δεν επιτρέπεται προτάσεις δοκιμής να υπάρχουν μέσα στα δεδομένα εκπαίδευσης!
Το μοντέλο θα αποδώσει μια εσφαλμένα μεγάλη πιθανότητα στις προτάσεις αυτές όταν θα τις δει στο σύνολο δοκιμής, και θα φανεί το μοντέλο καλύτερο από ότι είναι πραγματικά.

Χρειάζεται όμως στα πειράματα να κάνουμε πολλές δοκιμές, να κάνουμε αλλαγές που να οδηγούν σε καλύτερη απόδοση του μοντέλου, και αυτό μπορεί να έχει σαν αποτέλεσμα να προσαρμοστεί το μοντέλο στα δεδομένα δοκιμής, και να πάψει να είναι αντικειμενική η απόδοση στο σετ δοκιμής.
Η λύση: σετ επικύρωσης (validation/development set)
ένα τρίτο ξεχωριστό σετ δεδομένων που χρησιμοποιείται για την βελτιστοποίηση του μοντέλου μέχρι και το τελευταίο πείραμα. Και όταν έχουμε το τελικό μοντέλο, το εφαρμόζουμε μια και μοναδική φορά στο σετ δοκιμής που δεν το έχουμε ξαναδεί, και παίρνουμε την πραγματική απόδοση.

Perplexity: Ένα καλό ΜΓ είναι αυτό που δίνει μεγάλη πιθανότητα στο σετ δοκιμής

Όταν δυο ΜΓ εφαρμόζονται σε ένα test set, αυτό που δίνει την μεγαλύτερη πιθανότητα πρόβλεψης του test set είναι το καλύτερο ΜΓ.
Το καλύτερο ΜΓ είναι αυτό που 'εκπλήσσεται' λιγότερο από το test set. Είναι αυτό που θεωρεί το test set πιο αναμενόμενο.

Perplexity = η αντίστροφη πιθανότητα του test set, κανονικοποιημένη (διαιρεμένη) ως προς τον αριθμό των λέξεων του test set.

Γιατί κανονικοποίηση;
Η πιθανότητα του test set μικραίνει πολύ όσο μεγαλώνει το μέγεθός του. Άδικο: Δεν ισχύει ότι όσο μεγαλώνει ένα κείμενο σε μέγεθος, τόσο πιο λάθος είναι!
Χρειαζόμαστε οπότε ένα μέτρο που θα έχει αναχθεί σε επίπεδο λέξης.

$$PER\ (w_1, w_2, w_3, \ldots, w_N) = P(w_1, w_2, w_3, \ldots, w_N)^{-1/N}$$

Παίρνει τιμή [1, ∞], όσο πιο μεγάλη η τιμή, τόσο πιο μικρή η πιθανότητα που αποδίδεται στο test set, δηλαδή τόσο πιο λάθος, περίεργο, μη αναμενόμενο θεωρεί το ΜΓ το test set. Δηλ. τόσο πιο κακό είναι είναι το ΜΓ.

Αποτελέσματα

Training set: 38M λέξεις του Wall Street Journal Corpus
Test set: 1.5M λέξεις του Wall Street Journal Corpus

N-gram Order	Unigram	Bigram	Trigram
Perplexity	962	170	109

Πρόκληση: Sparse Data

Training set:
……έφαγε πρωινό
……έφαγε αυγό
……έφαγε το
……έφαγε βραδινό

Test set:
…..έφαγε πρωινό
…..έφαγε μεσημεριανό

P(μεσημεριανό | έφαγε) = 0
Άρα η πιθανότητα του test set είναι 0. Άρα το perplexity δεν υπολογίζεται (δεν μπορώ να διαιρέσω με 0).
Λύση: **Εξομάλυνση (Smoothing)**
Backoff (Υποχώρηση)
Παρεμβολή (Interpolation)

Add-1 ή Laplace Smoothing

Κάνουμε ότι βλέπουμε κάθε λέξη μια παραπάνω φορά από ό,τι την είδαμε στην πραγματικότητα.
Σε όλα τα counts προσθέτουμε 1.

Στα bigrams είχαμε $P(w_2|w_1) = count(w_1, w_2) / count(w_1)$
Με Add-1 Smoothing: $(P(w_2|w_1) = count(w_1, w_2) \mathbf{+ 1}) / (count(w_1) \mathbf{+ V})$,
όπου V μέγεθος του λεξιλογίου

Backoff & Interpolation

Η άλλη λύση όταν ακολουθίες λέξεων δεν εμφανίζονται στα δεδομένα εκπαίδευσης είναι να μειώσουμε το μέγεθος του παραθύρου συμφραζομένων που λαμβάνω υπόψη μου για την πρόβλεψη μιας λέξης.

Backoff: Αν έχω καλές ενδείξεις, χρησιμοποιώ τρίγραμμα, αλλιώς υποχωρώ σε δίγραμμα, αλλιώς σε μονόγραμμα.

Interpolation: Μίξη μονογράμμων, διγράμμων, τριγράμμων

$$\begin{aligned}\hat{P}(w_n|w_{n-2}w_{n-1}) &= \lambda_1 P(w_n|w_{n-2}w_{n-1}) \\ &\quad + \lambda_2 P(w_n|w_{n-1}) \\ &\quad + \lambda_3 P(w_n)\end{aligned}$$

$$\sum_i \lambda_i = 1$$

ΕΡΩΤΗΣΕΙΣ

1. Για να υπολογίσετε την πιθανότητα μιας πρότασης χρησιμοποιώντας μοντέλο διγράμμου

a) θα υπολογίσετε την δεσμευμένη πιθανότητα κάθε λέξης, δεδομένων όλων των λέξεων που προηγούνται αυτής, μέσα στην πρόταση, και θα πολλαπλασιάσετε τα αποτελέσματα
b) θα υπολογίσετε την δεσμευμένη πιθανότητα κάθε λέξης, δεδομένων των δυο λέξεων που προηγούνται αυτής, μέσα στην πρόταση, και θα πολλαπλασιάσετε τα αποτελέσματα
c) θα υπολογίσετε την δεσμευμένη πιθανότητα κάθε λέξης, δεδομένης της μιας λέξης που προηγείται αυτής, μέσα στην πρόταση, και θα πολλαπλασιάσετε τα αποτελέσματα
d) θα υπολογίσετε την δεσμευμένη πιθανότητα κάθε λέξης, χωρίς να λαμβάνετε υπόψη συμφραζόμενα αυτής, και θα πολλαπλασιάσετε τα αποτελέσματα

2. Έστω η παρακάτω πρόταση: *«Η IBM προσλαμβάνει μηχανικούς υπολογιστών»*. Στήστε τους τύπους που θα υπολογίσουν την πιθανότητα της πρότασης με μοντέλο γλώσσας

α. μονογράμμου β. διγράμμου γ. τριγράμμου

ΑΠΑΝΤΗΣΗ ΣΤΟ 2^{Ο} ΕΡΩΤΗΜΑ

«Η IBM προσλαμβάνει μηχανικούς υπολογιστών»

N=3 (τρίγραμμο)

$P(w_1, w_2, w_3, \ldots, w_n) \approx P(w_1) \times P(w_2|w_1) \times P(w_3|w_1, w_2) \times \ldots \times P(w_n| w_{n-2}, w_{n-1})$

P«Η IBM προσλαμβάνει μηχανικούς υπολογιστών»)=P(Η) x P(IBM|Η) x P(προσλαμβάνει |Η IBM) x P(μηχανικούς |IBM προσλαμβάνει) x P(υπολογιστών |προσλαμβάνει μηχανικούς) =

= count(Η)/μέγεθος του σώματος κειμένων

count(*Η IBM*) / count(*Η*) x

count(*Η IBM προσλαμβάνει*) / count(*Η IBM*) x

count(*IBM προσλαμβάνει μηχανικούς*) /count(*IBM προσλαμβάνει*) x

count(*προσλαμβάνει μηχανικούς υπολογιστών*) /count(*προσλαμβάνει μηχανικούς*)

ΣΥΝΕΧΕΙΑ

«Η IBM προσλαμβάνει μηχανικούς υπολογιστών»

N=2 (Bigram - δίγραμμο)

$P(w_1, w_2, w_3, \ldots, w_n) \approx P(w_1) \times P(w_2|w_1) \times P(w_3|w_2) \times \ldots \times P(w_n| w_{n-1})$

P«Η IBM προσλαμβάνει μηχανικούς υπολογιστών»)=P(Η) x P(IBM|Η) x P(προσλαμβάνει |IBM) x P(μηχανικούς |προσλαμβάνει) x P(υπολογιστών |μηχανικούς) =

= count(Η)/μέγεθος του σώματος κειμένων

count(*Η IBM*) / count(*Η*) x

count(*IBM προσλαμβάνει*) / count(*IBM*) x

count(*προσλαμβάνει μηχανικούς*) /count(*προσλαμβάνει*) x

count(*μηχανικούς υπολογιστών*) /count(*μηχανικούς*)

N=1 (Unigram – μονόγραμμο)

$P(w_1, w_2, w_3, \ldots, w_n) \approx P(w_1) \times P(w_2) \times P(w_3) \times \ldots \times P(w_n)$

P«Η IBM προσλαμβάνει μηχανικούς υπολογιστών»)=

P(Η) x P(IBM) x P(προσλαμβάνει) x P(μηχανικούς) x P(υπολογιστών)