Διαχείριση Web Περιεχομένου & Γλωσσικά Εργαλεία

Μάθημα 7ο

N-grams

Σοφία Στάμου

Άκ. Έτος 2008-09

Acknowledgement: McCoy, http://www.cis.udel.edu/~mccoy/courses/cisc882.03f/lectures/lect5-ngrams.ppt

Μια μική άσκηση

- Πώς θα συμπληρώνατε τις προτάσεις;
  - Ο διαιτητής ακύρωσε το .....
  - Σύνδεσμος, υπερσύνδεσμος, ποια η .....;
  - Όταν τελείωσα το διάβασμα πήγα .....

Πρόβλεψη λέξεων από τον άνθρωπο

- Κάποιοι προβλέπουν τις επόμενες λέξεις των προτάσεων
- Πώς;
  - Γνώση πεδίου
  - Γνώση σύνταξης
  - Γνώση λέξεων

Ισχυρισμός

- Μέρος της απαιτούμενης γνώσης για την πρόβλεψη των λέξεων εντοπίζεται με στατιστικά μοντέλα
- Δηλ. πιθανότητα μιας ακολουθίας (λέξεων, φράσεων)

Εφαρμογές

- Γιατί να προβλέψουμε την επόμενη λέξη βάσει των προηγούμενων;
  - Αυτόματη αναγνώριση λόγου, OCR, κτλ
  - Αξιολόγηση πρότασης για μηχανική μετάφραση, κτλ

Ορθογραφικά λάθη

- They are leaving in about fifteen *minuets* to go to her house.
- The study was conducted mainly *be* John Black.
- The design *an* construction of the system will take more than a year.
- Hopefully, all *with* continue smoothly in my absence.
- Can they *lave* him my messages?
- I need to *notified* the bank of….
- He is trying to *fine* out.

Παράδειγμα από Dorr, http://www.umiacs.umd.edu/~bonnie/courses/cmsc723-04/lecture-notes/Lecture5.ppt

Μοντελοποίηση γλώσσας

- Θεμελιώδες NLP εργαλείο
- Βασική ιδέα
  - Μερικές λέξεις συνεμφανίζονται πιο συχνά από άλλες
  - Η πιθανότητα συνεμφάνισης μπορεί να προβλεφθεί
  - Μπορούμε να υλοποιήσουμε ένα γλωσσικό μοντέλο

N-Grams

- Ακολουθίες από tokens
- N = πόσοι όροι εξετάζονται
  - Unigrams: 1 όρος
  - Bigrams: 2 όροι
  - Trigrams: 3 όροι
- Διαφορετικά είδη tokens
  - N-grams χαρακτήρων
  - N-grams λέξεων
  - N-grams Part of Speech
- Πληροφορία για το περιβάλλον συνεμφάνισης του token που εξετάζουμε

N-Gram μοντελοποίηση γλώσσας

- Γλωσσικό μοντέλο υπολογισμού πιθανότητας μιας πρότασης S, P(S).
- Το N-Gram μοντέλο χρησιμοποιεί τις προηγούμενες N-1 λέξεις σε μια ακολουθία για να προβλέψουν την επόμενη λέξη
- Πώς θα δημιουργήσουμε ή θα εκπαιδεύσουμε αυτά τα γλωσσικά μοντέλα;
  - Μέτρηση συχνοτήτων σε μεγάλα σώματα κειμένων
  - Καθορισμός πιθανοτήτων με χρήση μοντέλων Markov

Μετρώντας λέξεις σε σώματα κειμένων

- Τι είναι λέξη;

  δώρο και δώρα είναι μία λέξη;

  Σεπτέμβρης και Σεπτ.;

  Είναι λέξη το _, το ; , το (;

  Πόσες λέξεις είναι το εντωμεταξύ; Το παρόλη;

Ορολογία

- **Πρόταση**: μονάδα γραπτού λόγου
- **Εμφάνιση**: μονάδα προφορικού λόγου
- **Κλιτικός τύπος**: τύπος μιας λέξης στο σώμα κειμένων
- **Λήμμα**: ο πρώτος κλιτικός τύπος της λέξης
- **Τύποι**: ο αριθμός των διαφορετικών λέξεων του σώματος κειμένων (μέγεθος λεξιλογίου)
- **Tokens**: συνολικός αριθμός λέξεων

Απλά N-Grams

- Έστω μια γλώσσα με V τύπους λέξεων, ποια η πιθανότητα η λέξη x να ακολουθεί τη λέξη y;
  - Απλούστερο πιθανοτικό μοντέλο: 1/V
  - Εναλλακτικά (α): υπολογισμός της πιθανότητας το x να εμφανίζεται σε ένα νέο κείμενο βάσει της συχνότητας εμφάνισής του σε ένα σώμα κειμένων (unigram probability)
  - Εναλλακτικά (β): υπολογίζουμε τη δεσμευμένη πιθανότητα το x να εμφανίζεται στα συμφραζόμενα προηγούμενων λέξεων (bigrams, trigrams, …)

Υπολογισμός πιθανότητας

- Υπολογίζουμε το γινόμενο των επιμέρους δεσμευμένων πιθανοτήτων

  P(the mythical unicorn) = P(the) P(mythical|the) P(unicorn|the mythical)

- Όσο μεγαλύτερη η ακολουθία, τόσο μικρότερη η πιθανότητα να την εντοπίσουμε στο σώμα εκπαίδευσης

  P(Most biologists and folklore specialists believe that in fact the mythical unicorn horns derived from the narwhal)

- **Λύση**: υπολογίζουμε προσσεγιστικά με χρήση n-grams

Bigram μοντέλο

- Υπολογίσουμε προσεγγιστικά το $P(w_n|w_1^{n-1})$ από το $P(w_n|w_{n-1})$

- Υπόθεση Markov: η πιθανότητα μιας λέξης εξαρτάται μόνο από την πιθανότητα μιας περιορισμένης προγενέστερης γνώσης
- Γενικεύοντας: η πιθανότητα μιας λέξης εξαρτάται μόνο από την πιθανότητα των *n* προηγούμενων λέξεων
  - Trigrams, 4-grams, ….
  - Όσο μεγαλύτερο το *n* τόσο περισσότερα δεδομένα εκπαίδευσης απαιτούνται

Χρησιμοποιώντας N-grams

- Για τα μοντέλα N-Grams $P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1})$

  $P(w_{n-1},w_n) = P(w_n \mid w_{n-1}) P(w_{n-1})$

  Βάσει του **Αλυσιδωτού κανόνα** αποσυνθέτουμε στην από κοινού πιθανότητα (joint probability), π.χ. $P(w_1,w_2,w_3)$

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

  Για τα bigrams η πιθανότητα μιας ακολουθίας είναι το γινόμενο των δεσμευμένων πιθανοτήτων των bigrams που περιέχει

  P(the,mythical,unicorn) = P(unicorn|mythical) P(mythical|the) P(the|<start>)

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-1})$$

Παράδειγμα

P(I want to eat Chinese food) =

P(I | <start>) * P(want | I)
P(to | want) * P(eat | to)
P(Chinese | eat) * P(food | Chinese)

Μετρήσεις από το Berkeley Restaurant Project

Nth term

N-1 term	I	want	to	eat	Chinese	food	lunch
I	8	1087	0	13	0	0	0
want	3	0	786	0	6	8	6
to	3	0	10	860	3	0	12
eat	0	0	2	0	19	2	52
Chinese	2	0	0	0	0	120	1
food	19	0	17	0	0	0	0
lunch	4	0	0	0	0	1	0

Πίνακας Bigrams

Nth term

N-1 term

	I	want	to	eat	Chinese	food	lunch
I	.0023	.32	0	.0038	0	0	0
want	.0025	0	.65	0	.0049	.0066	.0049
to	.00092	0	.0031	.26	.00092	0	.0037
eat	0	0	.0021	0	.020	.0021	.055
Chinese	.0094	0	0	0	0	.56	.0047
food	.013	0	.011	0	0	0	0
lunch	.0087	0	0	0	0	.0022	0

Παράδειγμα

P(I want to eat Chinese food) =

P(I | <start>) * P(want | I)

P(to | want) * P(eat | to)

P(Chinese | eat) * P(food | Chinese)

- .25*.32*.65*.26*.02*.56 = .00015

Και;

- P(I want to eat British food) = P(I|<start>) P(want|I) P(to|want) P(eat|to) P(British|eat) P(food|British) = .25*.32*.65*.26*.001*.60 = .000080
- vs. I want to eat Chinese food = .00015
- Οι πιθανότητες εντοπίζουν *συντακτικά* γεγονότα, τη *γνώση για τον κόσμο*'
  - Το ρήμα eat ακολουθείται συχνά από NP
  - Το British food δεν είναι πολύ δημοφιλές

Τι μάθαμε για τη γλώσσα;

- Μετρήσαμε τα εξής:
  - P(want | I) = .32
  - P(to | want) = .65
  - P(eat | to) = .26
  - P(food | Chinese) = .56
  - P(lunch | eat) = .055

Και γι'αυτά;

- P(I | I) = .0023
- P(I | want) = .0025
- P(I | food) = .013

- Βρήκαμε ότι:
  - P(I | I) = .0023 I I I I want
  - P(I | want) = .0025 I want I want
  - P(I | food) = .013 the kind of food I want is ...
- Άρα, το σώμα κειμένων μας περιέχει λάθη

Παρατηρήσεις

- Ορισμένα γεγονότα έχουν μεγάλη συχνότητα εμφάνισης
- Πολλά γεγονότα έχουν μικρή συχνότητα εμφάνισης
- Μπορούμε να συλλέξουμε στατιστικά δεδομένα για πολύ συχνά γεγονότα εύκολα
- Ορισμένες μηδενικές συχνότητες γεγονότων είναι πράγματι μηδενικές, ενώ για άλλες τα γεγονότα απλά δεν έχουν εντοπιστεί ακόμα!
  - Πώς θα το αντιμετωπίσουμε;

Ο νόμος του Zipf

- Για πολλές κατανομές συχνοτήτων η *v*-ιοστή μεγαλύτερη συχνότητα είναι ανάλογη της αρνητικής δύναμης της σειράς ταξινόμησης του *v*
- Έστω *t* ανήκει σε ένα σύνολο μοναδικών γεγονότων. Ακόμα, έστω *f(t)* η συχνότητα του *t* και *r(t)* η σειρά ταξινόμησής του. Τότε:

$\forall t\ r(t) \approx c * f(t)^{-b}$ για κάποιες σταθερές b και c

Τεχνικές Smoothing

- Κάθε πίνακας εκπαίδευσης n-grams είναι αραιός για μεγάλα σώματα κειμένων (νόμος του Zipf)
- **Λύση**: υπολογισμός της πιθανότητας των n-grams που δεν έχουμε συναντήσει ακόμα
- **Προβλήματα**: πώς θα τροποποιήσουμε το υπόλοιπο σώμα κειμένου για να εντοπίσουμε αυτά τα *αόρατα* n-grams;
- Ένας τρόπος με smoothing τεχνικές

Add-one Smoothing

- Για unigrams:
  - Πρόσθεσε 1 σε κάθε count λέξης (τύπου)
  - Normalize με N (tokens) /(N (tokens) +V (types))
  - Smoothed count (προσαρμοσμένο για προσθήκες στο N):

$$\left(c_i+1\right)\frac{N}{N+V}$$

  - Normalize με N για τον υπολογισμό της πιθανότητας του νέου unigram:

$$\boxed{p_i^*=\frac{c_i+1}{N+V}}$$

- Για bigrams:
  - Πρόσθεσε 1 σε κάθε bigram c(wn-1 wn) + 1
  - Αύξησε το unigram count με το μέγεθος του λεξιλογίου V c(wn-1) + V

Witten-Bell Discounting

- Ένα μηδενικό n-gram είναι ένα unseen n-gram …όμως κάθε n-gram υπήρξε κάποτε unseen …συνεπώς...
  - Πόσες φορές συναντήσαμε ένα n-gram για πρώτη φορά; Μία για κάθε τύπο n-gram (T)
  - Υπολογίζουμε την πιθανότητα των unseen bigrams

$$\frac{T}{N+T}$$

  - Το σώμα εκπαίδευσης είναι μια ακολουθία γεγονότων, ένα για κάθε token (N) και ένα για κάθε νέο τύπο (T)
- Κατανέμουμε την πιθανότητα του συνόλου ισομερώς μεταξύ των unseen bigrams….ή υπολογίζουμε τη δεσμευμένη πιθανότητα ενός unseen bigram για την πρώτη λέξη του bigram

Backoff methods

- Για ένα trigram μοντέλο
  - Υπολογίζουμε τις πιθανότητες για το unigram,το bigram και το trigram
  - Στην πράξη:
    - Όταν δεν υπάρχει διαθέσιμο trigram υποχωρούμε στο bigram αν υπάρχει ή στην πιθανότητα του unigram

Συνοψίζοντας

- Οι N-gram πιθανότητες μπορούν να χρησιμοποιηθούν για να υπολογίσουμε την πιθανότητα
  - Μια λέξη να εμφανίζεται στο περιβάλλον (N-1)
  - Μια πρόταση να εμφανίζεται σε ένα κείμενο
- Οι τεχνικές smoothing αντιμετωπίζουν προβλήματα με λέξεις που δεν έχουμε συναντήσει σε ένα σώμα κειμένων
- Τα n-grams είναι χρήσιμα για μεγάλη ποικιλία NLP εργασιών

....στο επόμενο μάθημα

- Data mining

Ερωτήσεις…