MAI4CAREU

Επεξεργασία Φυσικής Γλώσσας

Language Modeling

Δημήτρης Πασχαλίδης

Τμήμα Επιστήμης Υπολογιστών

Πανεπιστήμιο Κύπρου

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

Τι είναι ένα Language Model?

- Υπολογισμός της πιθανότητας μιας πρότασης ή μιας ακολουθίας λέξεων.
  - $P(W) = P(w_1, w_2, w_3, \ldots, w_n)$
- Πιθανότητα επερχόμενης λέξης.
  - $P(w_5 \mid w_1, w_2, w_3, w_4)$
    - `Please turn off your cell _____`
    - `Your program does not _____`

→ Ένα μοντέλο ικανό να υπολογίσει είτε $P(W)$ ή $P(w_5 \mid w_1, w_2, w_3, w_4)$ ονομάζεται **Language Model**.

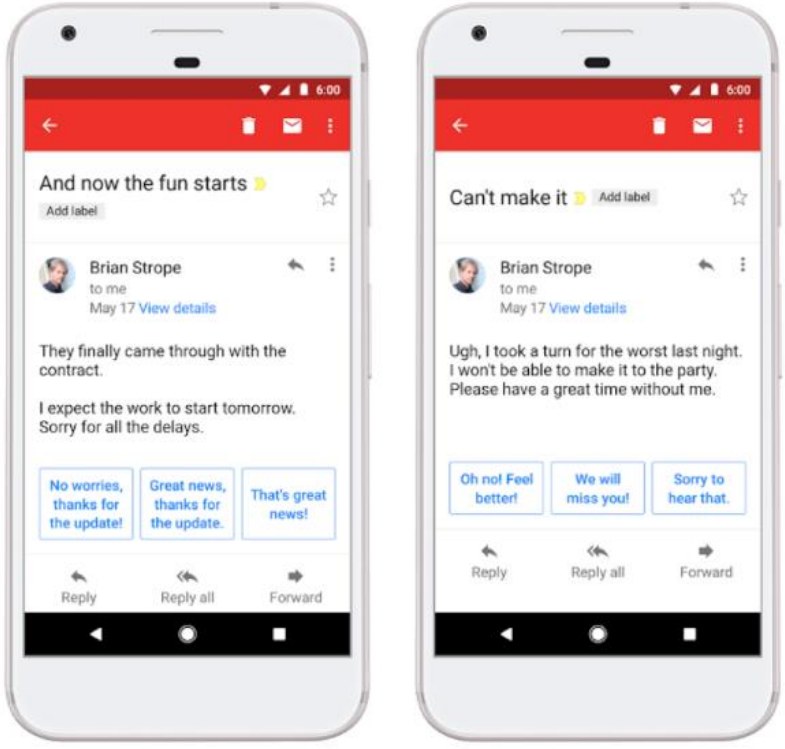
Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Εφαρμογή Language Model

- Machine Translation:
  - p(“*strong winds*”) > p(“*large winds*”)
- Sentence Completion:
  - P(“*Today is Tuesday*”) > P(“*Tuesday Today is*”)
- Spell Correction:
  - “*The office is about 15 minutes from my house*.”
  - p(“*15 minutes from my house*”) > p(“*15 minuets from my house*”)
- Speech Recognition
  - p(“*I saw a van*”) >> p(“*eyes awe of an*”)
- Summarization, question-answering, handwriting recognition, 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

Εφαρμογή Language Model



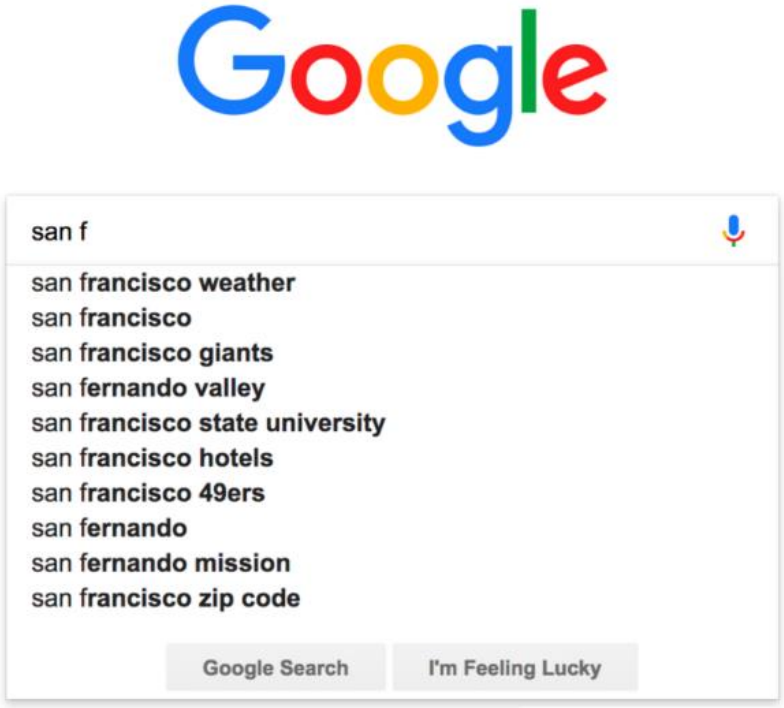


Image from Dr. Diyi Yang, CS 4650/7650 from Georgia Tech

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

Υπολογισμός Πιθανοτήτας P(W)

- Ας υπολογίσουμε την ακόλουθη κοινή πιθανότητα:
  - P(W) = P(*its*, *water*, *is*, *so*, *transparent*, *that*)
- Διαίσθηση: Βασιστείτε στον Αλυσιδωτό Κανόνα των Πιθανοτήτων:
  - Ορισμός υποθετικών πιθανοτήτων:
  - P(B | A) = P(A, B)/P(A) P(B | A)

$$P(x_1, x_2, x_3, \ldots, x_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

P(“*its water is so transparent*”) = P(*its*) x P(*water* | *its*) x P(*is* | *its water*) x P(*so* | *its water is*) x P(*transparent* | *its water is so*)

Πώς τα υπολογίζουμε;

Bag-of-Words με N-Grams

- N-grams: μια συνεχόμενη ακολουθία n tokens από ένα κείμενο.

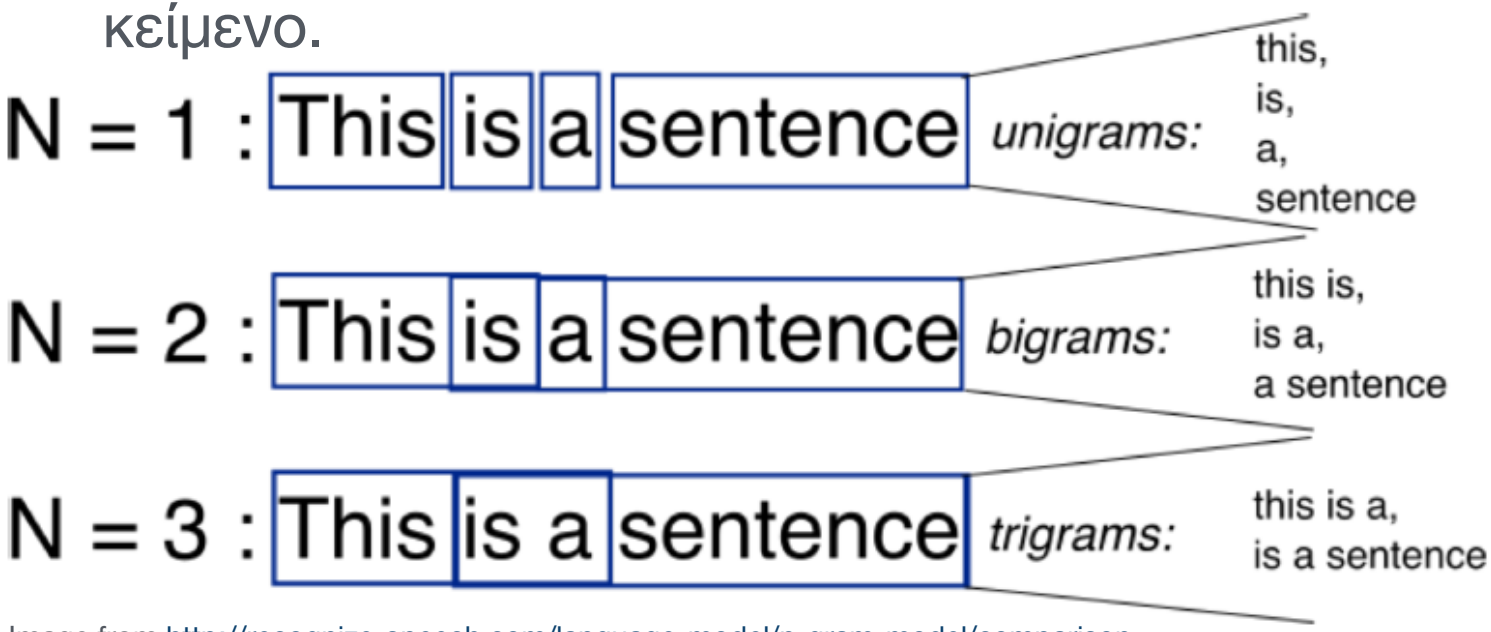


Image from http://recognize-speech.com/language-model/n-gram-model/comparison

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

N-Gram Models

- Unigram model: $P(w_1)P(w_2)P(w_3)\dots P(w_n)$
- Bigram model: $P(w_1)P(w_2|w_1)P(w_3|w_2)\dots P(w_n|w_{n-1})$
- Trigram model: $P(w_1)P(w_2|w_1)P(w_3|w_2,w_1)\dots P(w_n|w_{n-1}w_{n-2})$
- N-gram model: $P(w_1)P(w_2|w_1)\dots P(w_n|w_{n-1}w_{n-2}\dots w_{n-N})$

  - P(*the* | *its water is so transparent that*) = Count(*its water is so transparent that the*) / Count(*its water is so transparent that*)

Πάρα πολλές πιθανές προτάσεις και όχι αρκετά δεδομένα.

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

Markov Assumption

- Απλοποίηση της υπόθεσης:
  - P(the | its water is so transparent that) ≈ P(the | that)
- Ή ίσως:
  - P(the | its water is so transparent that) ≈ P(the | transparent that)
- Κατά προσέγγιση κάθε στοιχείο:
  - $P(w_i \mid w_1, w_2, \ldots, w_{i-1}) \approx P(w_i \mid w_{i-k}, w_{i-k-1}, \ldots, w_{i-1})$

Andrey Markov

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

Unigram Model k=1

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i)$$

- Παραδείγματα από ένα unigram model:

fifth, an, of, futures, the, an, incorporated, a,a, the, inflation, most, dollars, quarter, in, is, mass

thrift, did, eighty, said, hard, july, bullish

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

Bigram Model *k*=2

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx P(w_i \mid w_{i-1})$$

- Παραδείγματα από ένα bigram model:

`outside, new, car, parking, lot, of, the, agreement, reached`

`this, would, be, a, record, november`

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

Περιορισμοί

- Επέκταση σε trigrams, 4-grams, 5-grams etc.
- N-grams είναι γενικά ανεπαρκείς language models.
  - Η γλώσσα περιέχει εξαρτήσεις λέξεων μεγάλων αποστάσεων:

"*The computer I had just put on the machine room on the fifth floor crashed.*"

"*He is from France, so it makes sense that his first language is* _____"

- Αλλά συχνά μπορούμε να ξεφύγουμε με N-gram models.

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

Maximum Likelihood Estimate

- Εκτίμηση πιθανοτήτων bigram model :
  - $P(w_i \mid w_{i-1}) = count(w_{i-1}, w_i) \setminus count(w_{i-1})$
- Παράδειγμα:
  - <s> *I am Sam* <\s>
  - <s> *Sam I am* <\s>
  - <s> *I do not like green eggs and ham* <\s>

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

Πρακτικά Θέματα

- Κάνουμε τα πάντα στο *log* space:
  - Αποφυγή underflow
  - Η προσθήκη είναι ταχύτερη από τον πολλαπλασιασμό

$$\log(p_1 \times p_2) = \log(p_1) + \log(p_2)$$

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 από Restaurant Project

Reviews:

- Can you tell me about any good Cantonese restaurants close by.
- Mid priced thai food is what I'm looking for.
- Tell me about Chez Panisse.
- Can you give me a listing of the kinds of food that are available.
- I'm looking for a good place to eat breakfast.
- When is caffe Venezia open during the day.

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

Berkeley Restaurant Project

Reviews:

- Can you tell me about any good Cantonese restaurants close by.
- Mid priced thai food is what I'm looking for.
- Tell me about Chez Panisse.
- Can you give me a listing of the kinds of food that are available.
- I'm looking for a good place to eat breakfast.
- When is caffe Venezia open during the day.

Σύνολο 9222 προτάσεις

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

Ακατέργαστα Bigram Counts

	I	want	to	eat	Chinese	food	lunch	spend
I	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
Chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	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

Normalize διά Unigrams

	I	want	to	eat	Chinese	food	lunch	spend
Unigrams	**2533**	**927**	**2417**	**746**	**158**	**1093**	**341**	**278**
I	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
Chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	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

Υπολογισμός $P(w_i \mid w_{i-1})$

	I	want	to	eat	Chinese	food	lunch	spend
Unigrams	**2533**	**927**	**2417**	**746**	**158**	**1093**	**341**	**278**
I	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
Chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	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

Τι μάθαμε?

- P(*english*|*want*) = 0.0011
- P(*chinese*|*want*) = 0.0065

Στους ανθρώπους αρέσουν τα κινέζικα πράγματα περισσότερο στο specific corpus.

- P(*to*|*want*) = 0.66
- P(*eat* | *to*) = 0.28
- P(*food* | *to*) = 0.00
- P(*want* | *spend*) = 0.00
- P (*i* | <*s*>) = 0.25

Η αγγλική γλώσσα συμπεριφέρεται με έναν συγκεκριμένο τρόπο.

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

Παραγωγή γλώσσας

- Προσέγγιση Shakespeare:

1 gram	–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
2 gram	–Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow. –What means, sir. I confess she? then all sorts, he is trim, captain.
3 gram	–Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done. –This shall forbid it should be branded, if renown made it empty.
4 gram	–King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in; –It cannot be but so.

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

Παραγωγή γλώσσας

- Προσέγγιση Wall Street Journal:

1 gram	Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives
2 gram	Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her
3 gram	They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

- More: https://nbviewer.org/gist/yoavg/d76121dfde2618422139

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

Sparseness

- Maximum likelihood για την εκτίμηση του q.
  - Έστω $c(w_1, w_2, \ldots, w_n)$ είναι ο αριθμός των φορών που n-gram εμφανίζεται στο corpus.
  - Λεξιλόγιο από 20,000 λέξεις
    → αριθμός παραμέτρων είναι 8×10^{12}!

$$q(w_i|w_{i-2}, w_{i-1}) = \frac{c(w_{i-2}, w_{i-1}, w_i)}{c(w_{i-2}, w_{i-1})}$$

  - Τα περισσότερα n-grams δεν θα παρατηρηθεί ποτέ, ακόμη και αν είναι γλωσσικά εύλογο.
  - Οι περισσότερες προτάσεις θα έχουν μηδενικές ή απροσδιόριστες πιθανότητες

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

Αξιολόγηση Language Model

- **Extrinsic**: δημιουργήστε ένα νέο μοντέλο γλώσσας, χρησιμοποιήστε το για κάποια εργασία (MT, ASR, etc.)
- **Intrinsic**: μετρήστε πόσο καλοί είμαστε στη μοντελοποίηση της γλώσσας.

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

Extrinsic Αξιολόγηση

- Καλύτερη αξιολόγηση για τη σύγκριση των μοντέλων A και B
  - Τοποθέτηση κάθε μοντέλου σε μια εργασία: spelling corrector, speech recognizer, MT system
- Εκτέλεση της εργασίας → Αξιολόγηση του A και B
  - Πόσες λέξεις με ορθογραφικά λάθη διορθώθηκαν σωστά?
  - Πόσες λέξεις μεταφράστηκαν σωστά?
- Συγκρίνω accuracy για *A* και *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

Δυσκολία Extrinsic Αξιολόγησης

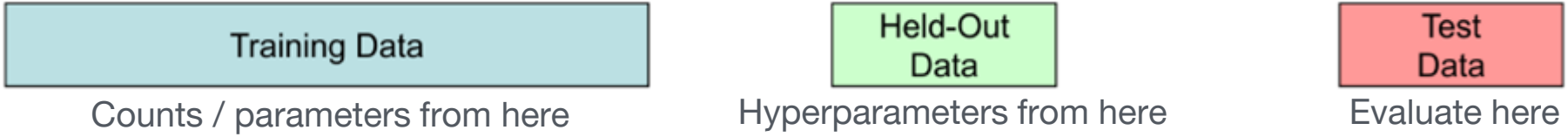
- **Extrinsic**: δημιουργήστε ένα νέο μοντέλο γλώσσας, χρησιμοποιήστε το για κάποια εργασία (MT, etc.)
  - Χρονοβόρα; μπορεί να διαρκέσουν ημέρες ή εβδομάδες
- Έτσι, μερικές φορές χρησιμοποιήστε **intrinsic evaluation**: perplexity
- Κακή προσέγγιση:
  - Εκτός αν test data μοιάζουν με training data
  - Έτσι γενικά χρήσιμο μόνο σε πιλοτικά πειράματα

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

Intrinsic Αξιολόγηση

- Διαισθητικά, τα γλωσσικά μοντέλα θα πρέπει να αποδίδουν μεγάλη πιθανότητα σε πραγματική γλώσσα που δεν έχουν ξαναδεί. Μεγιστοποιήστε την πιθανότητα επί test, όχι training data

- Τα μοντέλα απαιτούν τη ρύθμιση των παραμέτρων γενίκευσης σε δεδομένα αναμονής για την τόνωση της γενίκευσης των δοκιμών. Set hyperparameters για τη μεγιστοποίηση της πιθανότητας των διατηρούμενων δεδομένων.



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

Αξιολόγηση: Perplexity

- **Test data:** $S = \{s_1, s_2, \ldots, s_{sent}\}$
  - Οι παράμετροι δεν εκτιμώνται από S
  - Perplexity είναι η κανονικοποιημένη αντίστροφη

$$p(S) = \prod_{i=1}^{sent} p(s_i)$$

$$\log_2 p(S) = \sum_{i=1}^{sent} \log_2 p(s_i)$$

$$l = \frac{1}{M} \sum_{i=1}^{sent} \log_2 p(s_i)$$

$$\text{perplexity} = 2^{-l}$$

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

Αξιολόγηση: Perplexity

- **Perplexity:** $2^{-l}, l = \frac{1}{M}\sum_{i=1}^{sent} \log_2 p(s_i)$
- `sent` είναι ο αριθμός των προτάσεων στο test data.
- M είναι ο αριθμός των λέξεων στο test corpus

Ένα καλό LM έχει υψηλό p(S) και χαμηλό perplexity.

- **Training 38M λέξεις & Testing 1.5M λέξεις, WSJ**

N-Gram Order	Unigram	Bigram	Trigram
Perplexity	962	170	109

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

Ακατέργαστο Bigram Counts

	I	want	to	eat	Chinese	food	lunch	spend
I	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
Chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	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

Ακατέργαστες Bigram Πιθανότητες

	I	want	to	eat	Chinese	food	lunch	spend
Unigrams	**2533**	**927**	**2417**	**746**	**158**	**1093**	**341**	**278**
I	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
Chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	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

Language Generation

- Προσέγγιση Shakespeare:

1 gram	–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
2 gram	–Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow. –What means, sir. I confess she? then all sorts, he is trim, captain.
3 gram	–Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done. –This shall forbid it should be branded, if renown made it empty.
4 gram	–King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in; –It cannot be but so.

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

Shakespeare Corpus

- Tokens N = 884,647
- Ρήματα V = 29,066
- Παρήγαγε 300.000 bigrams από V^2=844M bigrams.
  - 99.96% των bigrams δεν είδαμε ποτέ (μηδενικές συμμετοχές).
- Quadrigrams χειρότερα: Η έξοδος μοιάζει Shakespeare γιατί στην πραγματικότητα είναι Shakespeare.

⇒ **Data Overfitting**

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

Data Overfitting

- N-grams models αποδίδουν καλά για την πρόβλεψη λέξεων εάν το test και training corpus είναι παρόμοια → Στην πραγματικότητα, αυτό συμβαίνει σπάνια.
- LMσ θα πρέπει να είναι **robust** και ικανά να **generalize**.
- **N-gram generalization**: Καταχωρήσεις με μηδενικές (0) εμφανίσεις.
  - Πράγματα που δεν είναι στο σετ προπόνησης αλλά είναι στο σετ δοκιμών.

Training set:
… denied the allegations
… denied the reports
… denied the claims
… denied the request

Test set:
… denied the offer
… denied the loan

P("offer" | denied the) = 0

Δεν είναι δυνατός ο υπολογισμός του perplexity (divide by 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

Smoothing

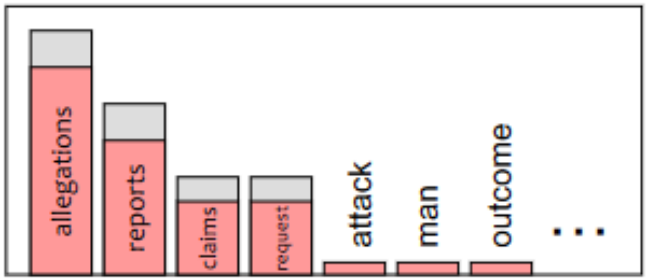
Η διαίσθηση του smoothing:

- Sparse statistics:

P(w | denied the)
3 allegations
2 reports
1 claims
1 request
7 total

- Steal probability mass to generalize:

P(w | denied the)
2.5 allegations
1.5 reports
0.5 claims
0.5 request
2 other
7 total



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

Laplace (Add-one) Smoothing

- Βασικά ένα **Add-one Estimation** σε όλες τις λέξεις/token.

- MLE estimate: $$P_{MLE}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

- Add-one estimate: $$P_{Add-1}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)+1}{c(w_{i-1})+V}$$

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

Add-one Smoothing Παράδειγμα

xya	100	100/300	**101**	**101/326**
xyb	0	0/300	**1**	**1/326**
xyc	0	0/300	**1**	**1/326**
xyd	200	200/300	**201**	**201/326**
xye	0	0/300	**1**	**1/326**
…				
xyz	0	0/300	1	1/326
Total xy	**300**	**300/300**	**326**	**326/326**

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

Berkeley Restaurant Corpus

	I	want	to	eat	Chinese	food	lunch	spend
I	6	828	1	10	1	1	1	3
want	3	1	609	2	7	7	6	2
to	3	1	5	687	3	1	7	212
eat	1	1	6	1	17	3	43	1
Chinese	2	1	1	1	1	83	2	1
food	16	1	16	1	2	5	1	1
lunch	3	1	1	1	1	2	1	1
spend	2	1	2	1	1	1	1	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

Laplace-Smoothed Bigrams

	I	want	to	eat	Chinese	food	lunch	spend
I	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
Chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.052	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

I	want	to	eat	Chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

$$P^*(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)+1}{C(w_{n-1})+V}$$

V = 1446 στο Berkeley Restaurant Project Corpus

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

Reconstruct Count Matrix

	I	want	to	eat	Chinese	food	lunch	spend
I	3.8	527	0.64	6.4	0.64	0.64	0.64	1.9
want	1.2	0.39	238	0.78	2.7	2.7	2.3	0.78
to	1.9	0.63	3.1	430	1.9	0.63	4.4	133
eat	0.34	0.34	1	0.34	5.8	1	15	0.34
Chinese	0.2	0.098	0.098	0.098	0.098	8.2	0.2	0.098
food	6.9	0.43	6.9	0.43	0.86	2.2	0.43	0.43
lunch	0.57	0.19	0.19	0.19	0.19	0.38	0.19	0.19
spend	0.32	0.16	0.32	0.16	0.16	0.16	0.16	0.16

$$C^*(w_{n-1}w_n) = P^*(w_n|w_{n-1}) \cdot C(w_{n-1}) = \frac{C(w_{n-1}w_n) + 1}{C(w_{n-1}) + V} \cdot C(w_{n-1})$$

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

Συγκρίνετε με τις πρώτες μετρήσεις

	I	want	to	eat	Chinese	food	lunch	spend
I	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
Chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0
I	3.8	527	0.64	6.4	0.64	0.64	0.64	1.9
want	1.2	0.39	238	0.78	2.7	2.7	2.3	0.78
to	1.9	0.63	3.1	430	1.9	0.63	4.4	133
eat	0.34	0.34	1	0.34	5.8	1	15	0.34
Chinese	0.2	0.098	0.098	0.098	0.098	8.2	0.2	0.098
food	6.9	0.43	6.9	0.43	0.86	2.2	0.43	0.43
lunch	0.57	0.19	0.19	0.19	0.19	0.38	0.19	0.19
spend	0.32	0.16	0.32	0.16	0.16	0.16	0.16	0.16

ced 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

Add-one Η εκτίμηση είναι ένα αμβλύ όργανο

- **Add-one Estimation** δεν χρησιμοποιείται για N-gram
- Ωστόσο, μπορεί να χρησιμοποιηθεί για την εξομάλυνση άλλων μοντέλων NLP:
  - Text classification models.
  - Τομείς όπου το μηδέν δεν είναι σημαντικό.

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

Backoff και Interpolation

- Ενίοτε **λιγότερο context είναι καλύτερο**.
- Backoff:
  - Χρήση trigram εάν έχετε βάσιμες αποδείξεις.
  - Αλλιώς bigram ή unigram
- Interpolation:
  - Mix unigram, bigram και trigram

Interpolation λειτουργεί καλύτερα.

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

Linear Interpolation

- Απλό Interpolation:

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

$$\sum_i \lambda_i = 1$$

- Lambdas Conditional επί Context:

$$\hat{P}(w_n|w_{n-2}w_{n-1}) = \lambda_1(w_{n-2}^{n-1}) P(w_n|w_{n-2}w_{n-1}) + \lambda_2(w_{n-2}^{n-1}) P(w_n|w_{n-1}) + \lambda_3(w_{n-2}^{n-1}) P(w_n)$$

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

Επιλογή Lambdas

- Χρησιμοποιήστε ένα held-out corpus

- Επιλογή λs για μεγιστοποίηση πιθανότητας held-out data:
  - Διορθώστε τις πιθανότητες n-gram (on training data).
  - Στη συνέχεια, αναζητήστε το λs που δίνουν τη μεγαλύτερη πιθανότητα στο held-out set:

$$\log P(w_1 \ldots w_n \mid M(\lambda_1 \ldots \lambda_k)) = \sum_i \log P_{M(\lambda_1 \ldots \lambda_k)}(w_i \mid w_{i-1})$$

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

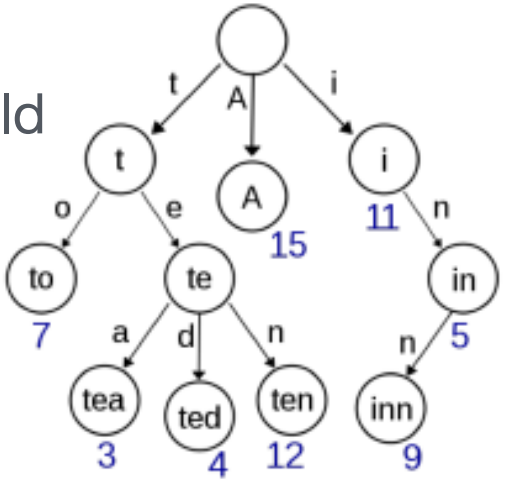
Out-of-Vocabulary Words OOV

- Κλειστό λεξιλόγιο vs. Ανοιχτό λεξιλόγιο
- Αντιμετωπίστε άγνωστες λέξεις:
  - Καλύψτε τέτοιους όρους με ένα ειδικό token <UNK>
  - Γλωσσικά μοντέλα επιπέδου χαρακτήρων

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

Web-scale N-Grams

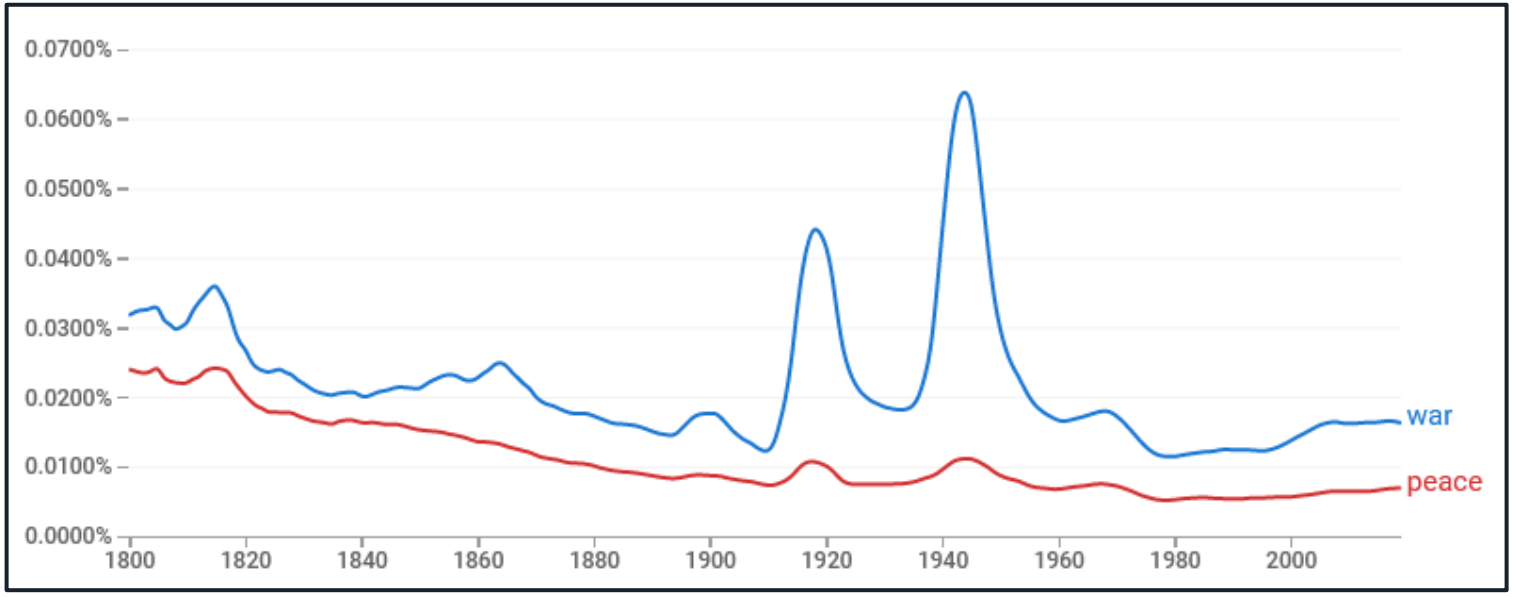
- Τεράστια n-gram corpus i.e. Google's N-Grams
- Pruning:
  - Λάβετε υπόψη N-Grams με μετρήσεις > threshold
  - Entropy-based pruning
- Efficiency:
  - Αποτελεσματικές δομές δεδομένων e.g. tries
  - Bloom filters: προσεγγιστικά language models
  - Αποθηκεύστε ως ευρετήρια π.χ.. Huffman coding



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

Google N-Grams

- Google n-gram viewer:
  https://books.google.com/ngrams/



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

- Language Models είναι ζωτικής σημασίας για το NLP. Υπολογίστε την πιθανότητα πρότασης/λέξης P(W)
- N-Gram Models: Uni-grams, Bi-grams, και Tri-grams
- Language Model Αξιολόγηση
  - Extrinsic
  - Intrinsic
  - Perplexity

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

Resources

- Jurafsky, D. and H. Martin Justin, Chapter 3. "N-gram Language Models" Speech and Language Processing
- Python Tutorial on N-Grams using NLTK: https://www.askpython.com/python/examples/n-grams-python-nltk
- Google N-Gram Viewer: https://books.google.com/ngrams

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