MAI4CAREU

Επεξεργασία Φυσικής Γλώσσας
Βασική προ-επεξεργασία κειμένου

Δημήτρης Πασχαλίδης

Τμήμα Πληροφορικής

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

Regular Expressions RegEx

Μια επίσημη γλώσσα για τον καθορισμό και την αναζήτηση συμβολοσειρών κειμένου.
Πώς μπορούμε να αναζητήσουμε οποιοδήποτε από αυτά?

- Woodchuck
- Woodchucks
- woodchuck
- woodchucks

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

RegEx: Disjunctions

Γράμματα μέσα σε αγκύλες [ ]

Pattern	Matches
`[wW]oodchuck`	**W**oodchuck, **w**oodchuck
`[1234567890]`	Any digit

Χρησιμοποιήστε εύρη τιμών [A-Z]

Pattern	Matches
`[A-Z] ή [a-z]`	**D**renched **b**lossoms
`[0-9]`	Chapter **1**: Down the Rabbit Hole

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

RegEx: Negation

Αρνήσεις [^Ss]

Pattern	Matches	
[^A-Z]	Όχι κεφαλαίο γράμμα	W**o**odchuck, woodchuck
[^Ss]	Ούτε ‘*S*’ ούτε ‘*s*’	**I** have no exquisite reason”
[^e^]	Ούτε e ούτε ^	Look h**e**re
a^b	Βρείτε αντιστοιχίσεις για “*a^b*”	Look up **a^b** now

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

RegEx: ? * + .

Pattern	Matches	
colou?r	Προαιρετικός προηγούμενος χαρακτήρας	**color**, **colour**
oo*h!	0 ή μέρος του προηγούμενου χαρακτήρα	**oh!** **ooh!** **oooh!** **ooooh!**
oo+h!	1 ή περισσότεροι προηγούμενοι χαρακτήρες	~~**oh!**~~ **ooh!** **oooh!** **ooooh!**
baa+		**baa**, **baaa**, **baaaaaa**
beg.n		**begin**, **begun**, **beg3n**

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

RegEx: More Disjunctions

Μπορούμε να καθορίσουμε σύνθετες κανονικές εκφράσεις ενώνοντας ξεχωριστά regexes με έναν τελεστή διάζευξης |

Pattern	Matches
`waters?|earth`	`as if the **waters** had but newly retired from the face of the **earth**,`

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

RegEx: Example

Εύρεση όλων των εμφανίσεων της λέξης "the" σε κείμενο.

Pattern	Matches
the	Χάνει όλα τα κεφαλαία *The* ή *THE*
[tT]he	Επιστρέφει εσφαλμένα "theology" ή "isothermally"
[^a-zA-Z][tT]he[^a-zA-Z]	Σωστό
/\b[Tt]he\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

RegEx: Symbols

Symbol	Function
\b	Όριο λέξης (μηδενικό πλάτος)
\d	Οποιοδήποτε δεκαδικό ψηφίο = [0-9]
\D	Οποιοσδήποτε μη ψηφιοποιημένος χαρακτήρας = [^0-9]
\s	Οποιοσδήποτε χαρακτήρας κενού διαστήματος = [\t\n\r\f\v]
\S	Οποιοσδήποτε χαρακτήρας χωρίς κενό διάστημα = [^ \t\n\r\f\v]
\w	Οποιοσδήποτε αλφαριθμητικός χαρακτήρας = [a-zA-Z0-9_]
\W	Οποιοσδήποτε μη αλφαριθμητικός χαρακτήρας = [^a-zA-Z0-9_]

Taken from https://www.nltk.org/book/ch03.html

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

Regular expressions έχουν κρίσιμο ρόλο στο NLP

- Οι ακολουθίες του regex είναι συχνά
  το πρώτο μοντέλο για οποιοδήποτε κείμενο επεξεργασίας κειμένου.
  Στη Μηχανική Μάθηση τα regex εξακολουθούν να χρησιμοποιούνται για προ-επεξεργασία, ή ως features στους classifiers
  - Μπορεί να είναι πολύ χρήσιμα στις γενικεύσεις

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

Προ-επεξεργασία κειμένου

Κάθε εργο NLP απαιτεί προ-επεξεργασία κειμένου:

1. Tokenization των λέξεων
2. Κανονικοποίηση λέξεων
   Segmentation προτάσεων

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

Tokenization Λέξεων

Tokenization βάσει διαστήματος (space): `text.split(" ")`

```
As much mud in the streets as if the waters
had but newly retired from the face of the
earth, and it would not be wonderful to meet a
Megalosaurus, forty feet long or so, waddling
like an elephantine lizard up Holborn Hill.
```

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

Tokenization Λέξεων

Tokenization βάσει διαστήματος (space): `text.split(" ")`

```
As much mud in the streets as if the waters had but newly retired from the face of the earth, and it would not be wonderful to meet a Megalosaurus, forty feet long or so, waddling like an elephantine lizard up Holborn Hill.
```

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

Tokenization Λέξεων

Tokenization βάσει διαστήματος (space): `text.split(" ")`

```
As much mud in the streets as if the waters had but newly retired from the face of the earth, and it would not be wonderful to meet a Megalosaurus, forty feet long or so, waddling like an elephantine lizard up Holborn Hill.
```

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

Tokenization Λέξεων

Εμφανήσεις του "*earth*" σε 100 βιβλία (Project Gutenberg)

earth	368
earth,	135
earth.	68
earth.”	26
earth’s	24
earth!	18
earth;	16

Issues in Tokenization

Cannot blindly remove punctuation:

- Ph.D., AT&T, isn’t
- $45.55
- 01/02/2022
- http://www.google.com
- #HaveANiceDay
- dpasch01@cs.ucy.ac.cy
- ;P or :) or :’(

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

Tokenization Λέξεων

```
text = "That U.S.A. poster-print costs $12.40..."

pattern = r""" (?x)                  # set flag to allow verbose regexps
               (?:[A-Z]\.)+          # abbreviations, e.g. U.S.A.
               |\d+(?:\.\d+)?%?      # numbers, incl. currency and percentages
               |\w+(?:[-']\w+)*      # words w/ optional internal hyphens
               |(?:[+/\-@&*])        # special characters with meanings
           """

  print('', nltk.regexp_tokenize(text, pattern))
```

```
 ['That', 'U.S.A.', 'poster-print', 'costs', '12.40']
```

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

Tokenization Λέξεων- Data-oriented Προσεγγιση

- Αντί για κενό διάστημα → Tokenize βάσει δεδομένων
- Τρεις κοινοί αλγόριθμοι:
  - Byte-Pair Encoding (BPE) (Sennrich et al. 2016)
  - Unigram language modeling tokenization (Kudo, 2018)
  - WordPiece (Schuster and Nakajima, 2012)
- Μοτίβο Μεθοδολογίας:
  1. Token learner αναγνωρίζει το λεξιλόγιο μέσα από το raw training corpus (σύνολο από tokens)
  2. Ο token segmenter λαμβάνει τις προτάσεις και τις κάνει tokenize βάσει του λεξιλογίου.

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

Byte Pair Encoding (BPE) Token Learner

- Αφήστε το λεξιλόγιο να είναι ένα σύνολο όλων των μεμονωμένων χαρακτήρων:

  `= {A, B, C, D, …, a, b, c, d, …}`

- Επαναλάβετε τα εξής:
  1. Επιλέξτε τα δύο σύμβολα που γειτνιάζουν συχνότερα στο training corpus (e.g. `'A'` and `'B'`)
  2. Προσθήκη νέου συμβόλου `'AB'` στο λεξιλόγιο, που προέρχεται από `'A'` και `'B'`.
  3. Αντικαταστήστε κάθε παρακείμενο `'A'` και `'B'` στο corpus με `'AB'`

- Μέχρι k συγχωνεύσεις.

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

Byte Pair Encoding (BPE) Token Learner

- Αρχικό corpus:

  `low low low low low lowest lowest newer newer newer newer newer newer wider wider wider new new`

- Προσθεσε τα tokens που αντιστοιχούν στα τέλη λέξεων και δημιουργήστε το λεξιλόγιο:

{ _, d, e, i, l, n, o, r, s, t, 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

Byte Pair Encoding (BPE) Token Learner

- Αρχικό corpus και λεξιλόγιο:

Token	Freq.	Vocabulary
l o w _	5	_, d, e, i, l, n, o, r, s, t, w
l o w e s t _	2	
n e w e r _	6	
w i d e r _	3	
n e w _	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

Byte Pair Encoding (BPE) Token Learner

- Merge `e r` προς `er`

Token	Freq.	Vocabulary
`l o w _`	5	`_, d, e, i, l, n, o, r, s, t, w, er`
`l o w e s t _`	2	
`n e w er _`	6	
`w i d er _`	3	
`n e w _`	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

Byte Pair Encoding (BPE) Token Learner

- Merge `er _` προς `er_`

Token	Freq.	Vocabulary
`l o w _`	5	`_, d, e, i, l, n, o, r, s, t, w, er, er_`
`l o w e s t _`	2	
`n e w er_`	6	
`w i d er_`	3	
`n e w _`	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

Byte Pair Encoding (BPE) Token Learner

❑ Merge `n e` προς `ne`

Token	Freq.	Vocabulary
`l o w _`	5	`_, d, e, i, l, n, o, r, s, t, w, er, er_, ne`
`l o w e s t _`	2	
`ne w er_`	6	
`w i d er_`	3	
`ne w _`	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

Byte Pair Encoding (BPE) Token Learner

- Τα υπόλοιπα merges είναι:

Merge	Vocabulary
ne, w	_, d, e, i, l, n, o, r, s, t, w, er, er , ne, new
l, o	_, d, e, i, l, n, o, r, s, t, w, er, er , ne, new, lo
lo, w	_, d, e, i, l, n, o, r, s, t, w, er, er , ne, new, lo, low
new, er_	_, d, e, i, l, n, o, r, s, t, w, er, er , ne, new, lo, low, newer_
low, _	_, d, e, i, l, n, o, r, s, t, w, er, er , ne, new, lo, low, newer_, low_

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

Byte Pair Encoding (BPE) Token Segmenter

- Στα δεδομένα δοκιμής, εκτελέστε κάθε merge:
- Αρχικά, merge κάθε `e r` προς `er`, τότε `er _` προς `er_` etc.
- Αποτέλεσμα:
  - Σετ δοκιμής `n e w e r _` θα γινόταν tokenized ως πλήρης λέξη
    Σετ δοκιμής `l o w e r _` θα ήταν δύο tokens: `low` και `er_`

BPE είναι σε θέση να συλλάβει μορφήματα όπως -est ή -er

Ένα μόρφημα είναι η μικρότερη μονάδα που φέρει νόημα μιας γλώσσας
Παράδειγμα `unlikeliest` έχει 3 **μορφήματα** `un-`, `likely`, και `-est`.

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

Κανονικοποίηση λέξεων

- Η τοποθέτηση λέξεων/διακριτικών σε τυποποιημένη μορφή
- Dimensionality reduction του λεξιλογίου του corpus:
  - *U.S.A* ή *USA*
  - *uhhuh* ή *uh-huh*
  - *Fed* ή *fed*
  - *am*, *is*, *be*, *are*

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

Κανονικοποίηση λέξεων: Case Folding

- Μείωση όλων των γραμμάτων σε πεζά γράμματα:
  - Οι χρήστες τείνουν να χρησιμοποιούν πεζά
    Πιθανές εξαιρέσεις:
    - General Motors
    - Fed vs. fed
    - SAIL vs. sail
  - Χρήσιμο για ανάλυση συναισθήματος και εξαγωγή πληροφοριών
    Θέματα: US vs. us

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

Κανονικοποίηση λέξεων: Lemmatization

- Αναπαραστήστε όλες τις λέξεις ως λήμμα τους:
  - `am, are, is → be`
  - `car, cars, car's, cars' → car`
  - `"He is reading detective stories"`
    `→ "He be read detective story"`
- Lemmatization γίνεται από **Morphological Parsing**
  - Parse token σε μορφώματα → `cars'`
  - **Stems**: Βασικές μονάδες που φέρουν νόημα → `car`
  - **Affixes**: Άλλα μέρη με γραμματικές λειτουργίες → `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

Κανονικοποίηση λέξεων: Stemming

- Ευρετική διαδικασία για την αφαίρεση των inflected suffixes ενος token:
  - `organizes, organized, organizing` → `organ`
  - Χαμηλότερη precision, υψηλότερη recall
- Δημοφιλής **Porter Stemmer**
  - Μια ακολουθία κανόνων για την αφαίρεση suffixes
    Παραδείγματα:
    - ATIONAL → ATE (e.g. `relational` → `relate`)
    - ING → ϵ αν stem περιέχει φωνήεν (e.g. `motoring` → `motor`)
    - SSNES → SS (e.g. `grasses` → `grass`)

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

Segmentation Πρότασεων

- Tokenization λέξεων είναι ένα στάδιο προ-επεξεργασίας του sentence segmentation → Προσδιορισμός των ορίων μεταξύ των προτάσεων.
- Αρκετές εργασίες NLP λειτουργούν σε επίπεδο πρότασης (e.g. POS tagging, parsing) → Sentence segmentation είναι σημαντικό
- Pattern: Tokenize πρώτα χρησιμοποιώντας κανόνες ή ML για να classify περιόδους ως
  - Μέρος ενός token or
  - Ένα όριο πρότασης

- "Do you want to go?" said Jane.
- Mr. Collins said he was going.
- He lives in the U.S. John, however, lives in Canada

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

- Tokenization λέξεων, κανονικοποίηση, και segmentation προτάσεων θεωρούνται βασικές αρχές στην προ-επεξεργασία κειμένου.
- Tokenization λέξεων όχι απλά διαχωρισμός λευκού χώρου αλλά με τη χρήση δεδομένων e.g. *Byte Pair Encoding (BPE) algorithm*.
- Κανονικοποίηση λέξεων: *lemmatization* and *stemming*.

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

Text Similarity

- Οι άνθρωποι μπορούν να εκφράσουν την ίδια έννοια (ή σχετικές έννοιες) με πολλούς διαφορετικούς τρόπους.
  - Παράδειγμα "*the plane leaves at 12pm*" vs. "*the flight departs at noon*".
- Text similarity αποτελεί βασικό συστατικό του NLP.
- Πληροφορίες σχετικά με "*cats*" → θέλουμε επίσης αναφορές σε "*kittens*" (λέξη "*cat*" δεν είναι σε αυτό).
- Πληροφορίες σχετικά με "*fruit dessert*" → θέλουμε επίσης "*peach tart*" και "*apple cobbler*".

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

Ανθρώπινα Annotations για Similarity

- Annotators ερωτήθηκαν πόσο παρόμοιες είναι δύο λέξεις; **Maximum similarity score είναι 10.**

Finkelstein, L., Gabrilovich, et al. 2002 Placing search in context: The concept revisited. WWW 2002 (pp. 406-414).

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

Ανθρώπινα Annotations για Similarity

- Χρησιμοποιείται επίσης Part-of-Speech ετικέτες.
- **“*Persuade*” μοιάζει περισσότερο με “*argue*” παρά με “*pursue*”.**

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

Τύποι Text Similarity

- Morphological similarity (e.g. *respect* - *respectful*)
- Spelling similarity (e.g. *theater* - *theatre*)
- Synonym (e.g. *car* - *automobile*)
- Semantic similarity (e.g. *Spain* - *Barcelona*)

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

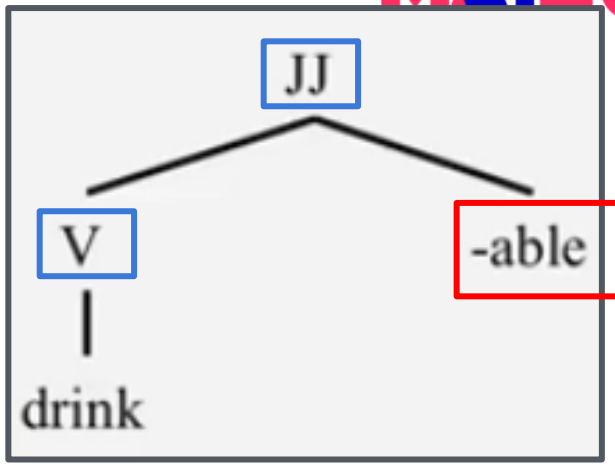
Τύποι Text Similarity

- Morphological similarity (e.g. *respect* - *respectful*)
- Spelling similarity (e.g. *theater* - *theatre*)
- Synonym (e.g. *car* - *automobile*)
- Semantic similarity (e.g. *Spain* - *Barcelona*)

Word embeddings → Περισσότερα για αυτό στη διάλεξη 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

Morphological Similarity



Morpheme (slide 23)

Part-of-Speech Tags
- JJ = Adjective
- V = Verb

- Derivational Morphology
  - Παράδειγμα: "*drinkable*"
  - Morphemes: *-able, -ing, -re, -un, -er*
- Λέξεις με την ίδια ρίζα:
  - Βασική μορφη: *scan*
  - Inflected μορφες: *scans*, *scanned*, *scanning*
  - Suffixes: *scanner*
  - Prefixes: *rescan*

Morphological similarity → Χρήση του **Stemming** (slide #)

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

Spelling Similarity

- Λέξεις με typos και spelling variants:
  - Reciept → receipt
  - Catherine → Katherine (αντικατάσταση of ‘C’ with ‘K’)
  - Polarization → polarisation (αντικατάσταση of ‘z’ with ‘s’)
  - Theater → theatre
- Edit Operations:
  - Insertion/Deletion
  - Substitution

} Levenshtein Distance (ή Minimum Edit Distance)

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

Minimum Edit Distance (MED)

- Βασισμένο σε δυναμικό προγραμματισμό Insertions, deletions, και substitutions συνήθως όλα έχουν κόστος 1.
- Παράδειγμα από similarity ανάμεσα s_1= “*strength*” και s_2 = “*trend*”

		s	t	r	e	n	g	t	h
	0								
t									
r									
e									
n									
d									

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

Recurrence Relation

Ορισμοί:

- $s_1(i)$: i^{th} character in s_1.
- $s_2(j)$: j^{th} character in s_2.
- D(i, j): edit distance ανάμεσα s_1 prefix μήκους i και s_2 prefix μήκους j.
- t(i, j): κόστος ευθυγράμμισης του i^{th} χαρακτήρα σε string s_1 με j^{th} character in string s_2.

Αναδρομικές εξαρτήσεις:

- D(i, 0) = i
- D(0, j) = j
- D(i, j) = min([
  D(i-1, j) + 1,
  D(i, j-1) + 1,
  D(i-1, j-1) + t(i, j)
  ])
- t(i, j) = 0 iff s1(i) = s2(j) else t(i, j)=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

MED Παράδειγμα

- Υπολογίζω edit distance of ‘s’:
  - D(i, 0) = i

		s	t	r	e	n	g	t	h
	0	1							
t									
r									
e									
n									
d									

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

MED Παράδειγμα

- Υπολογίζω edit distance of '*s*':
  - D(i, 0) = i
- Υπολογίζω edit distance of '*t*':
  - D(0, j) = j

		s	t	r	e	n	g	t	h
	0	1							
t	1								
r									
e									
n									
d									

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

MED Παράδειγμα

- Υπολογίζω edit distance of 's':
  - D(i, 0) = i
- Υπολογίζω edit distance of 't':
  - D(0, j) = j
- Κάντε το ίδιο με όλους i^{th} και j^{th} γράμματα απο s_1 και s_2.

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1								
r	2								
e	3								
n	4								
d	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

MED Παράδειγμα

- Υπολογίζω edit distance between ‘s’ και ‘t’:
  - **D(i, j)** = min([
    D(i - 1, j) + 1,
    D(i, j - 1) + 1,
    D(i - 1, j - 1) + t(i, j)
    ])

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1								
r	2								
e	3								
n	4								
d	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

MED Παράδειγμα

- Υπολογίζω edit distance ανάμεσα ‘s’ και ‘t’:
  - **D(1, 1)** = min([
    D(1 - 1, 1) **+ 1,**
    D(1, 1 - 1) **+ 1,**
    D(1 - 1, 1 - 1) **+ t(1, 1)**
  ])

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1								
r	2								
e	3								
n	4								
d	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

MED Παράδειγμα

- Υπολογίζω edit distance ανάμεσα ‘s’ και ‘t’:
  - **D(1, 1)** = min([
    - **1** + 1,
    - **1** + 1,
    - **0** + t(1, 1)

  ])

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1								
r	2								
e	3								
n	4								
d	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

MED Παράδειγμα

- Υπολογίζω edit distance ανάμεσα ‘s’ και ‘t’:
  - **D(1, 1)** = min([
    1 + 1,
    1 + 1,
    0 + 1
    ])

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1								
r	2								
e	3								
n	4								
d	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

MED Παράδειγμα

- Υπολογίζω edit distance ανάμεσα ‘s’ και ‘t’:
  - **D(1, 1)** = min([
    2,
    2,
    1
    ])

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1								
r	2								
e	3								
n	4								
d	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

MED Παράδειγμα

- Υπολογίζω edit distance ανάμεσα ‘s’ και ‘t’:
  - **D(1, 1)** = 1

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1							
r	2								
e	3								
n	4								
d	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

MED Παράδειγμα

- Υπολογίζω edit distance ανάμεσα ‘t’ και ‘t’:
  - **D(2, 1)** = min([
    D(1, 1) **+ 1,**
    D(2, 0) **+ 1,**
    D(1, 0) **+ t(i, j)**
    ])

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1							
r	2								
e	3								
n	4								
d	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

MED Παράδειγμα

- Υπολογίζω edit distance ανάμεσα ‘t’ και ‘t’:
  - **D(2, 1)** = min([
    1 + 1,
    2 + 1,
    1 + t(i, j)
  ])

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1							
r	2								
e	3								
n	4								
d	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

MED Παράδειγμα

- Υπολογίζω edit distance ανάμεσα ‘t’ και ‘t’:
  - **D(2, 1)** = min([
    2,
    3,
    1 + t(i, j) = 0
    ])

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1							
r	2								
e	3								
n	4								
d	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

MED Παράδειγμα

- Υπολογίζω edit distance ανάμεσα ‘t’ και ‘t’:
  - **D(2, 1)** = min([
    2,
    3,
    1
    ])

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1							
r	2								
e	3								
n	4								
d	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

MED Παράδειγμα

- Υπολογίζω edit distance ανάμεσα ‘t’ και ‘t’:
  - **D(2, 1)** = 1

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1						
r	2								
e	3								
n	4								
d	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

MED Παράδειγμα

- Συνεχίστε τη διαδικασία για τα ακόλουθα cells.

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2						
e									
n									
d									

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

MED Παράδειγμα

- Συνεχίστε τη διαδικασία για τα ακόλουθα cells.
- Υπολογίζω edit distance ανάμεσα ‘t’ και ‘t’:
  - **D(3, 2)** =

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2						
e									
n									
d									

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

MED Παράδειγμα

- Συνεχίστε τη διαδικασία για τα ακόλουθα cells.
- Υπολογίζω edit distance ανάμεσα ‘t’ και ‘t’:
  - **D(3, 2)** = 1

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2	1					
e									
n									
d									

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

MED Παράδειγμα

- Ολοκληρώστε τη διαδικασία για τα ακόλουθα cells.

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2	1	2	3	4	5	6
e	3	3	3	2	1	2	3	4	5
n	4	4	4	3	2	1	2	3	4
d	5	5	5	4	3	2	2	3	4

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

MED Παράδειγμα

- Ολοκληρώστε τη διαδικασία για τα ακόλουθα cells.
- Ο πίνακας ονομάζεται Edit Transcript.
- The MED διαδρομή είναι highlighted.
- Η τελική καταχώρηση στο MED path είναι η minimum edit distance ανάμεσα “*strength*” και “*trend*”

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2	1	2	3	4	5	6
e	3	3	3	2	1	2	3	4	5
n	4	4	4	3	2	1	2	3	4
d	5	5	5	4	3	2	2	3	**4**

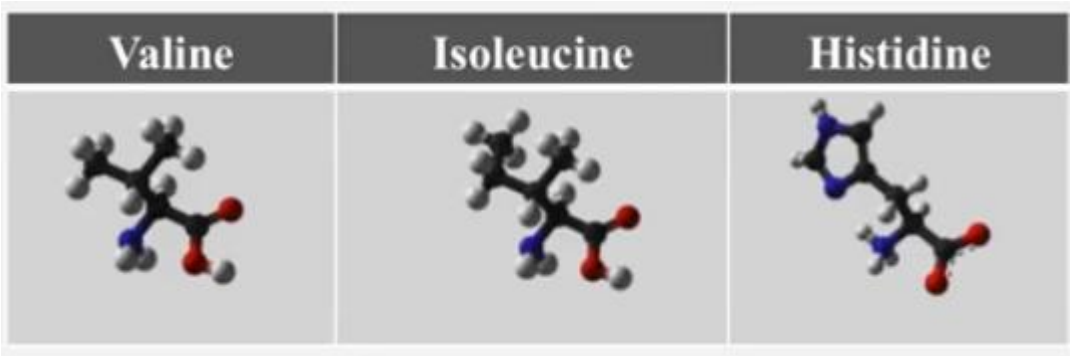
Useful online tool for MED calculation: https://planetcalc.com/1721/

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

Άλλες εφαρμογές του MED

AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAACCTCCCATCCGTGTCTA
TTGTACCCTGTTGCTTCGGCGGGCCCGCCGCTTGTCGGCCGCCGGGGGGGCGCCTCTGCCCC
CCGGGCCCGTGCCCGCCGGAGACCCCAACACGAACACTGTCTGAAAGCGTGCAGTCTGAGTT
GATTGAATGCAATCAGTTAAAACTTTCAACAATGGATCTCTTGGTTCCGGC

- Γενετική αλληλουχία (nucleotides AGCT)
  → Computational Biology
- Στοίχιση μη κειμενικών ακολουθιών



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

- Text similarity αποτελεί βασικό συστατικό του NLP.
- Τύποι text similarity συμπεριλαμβάνουν:
  - Morphological similarity (e.g. respect - respectful)
  - Spelling similarity (e.g. theater - theatre)
  - Synonym (e.g. car - automobile)
  - Semantic similarity (e.g. Spain - Barcelona)
- Minimum Edit (ή Levensteing) distance είναι θεμελιώδης στο 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

Πηγές

- Jurafsky, D. and H. Martin Justin, Chapter 2. "Regular Expressions, Text Normalization, Edit Distance" Speech and Language Processing
- Python Regex How To: https://docs.python.org/3/howto/regex.html
- Natural Language Toolkit: https://www.nltk.org/index.html

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