ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ · ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Γνωριμία με το NLTK

Γλωσσική Τεχνολογία ·

Δέσποινα Μουρατίδη

Τι θα κάνουμε σήμερα

1

Google Colab

Python στον browser, χωρίς εγκατάσταση

2

Κείμενα

Εννέα αγγλικά έργα του nltk.book

3

Μετρήσεις

Tokens, types, λεξιλογική ποικιλία

4

Κανονικές εκφράσεις

Αναζήτηση προτύπων με το re_show()

Τι είναι το NLTK

Natural Language Toolkit: βιβλιοθήκη Python ανοιχτού κώδικα για επεξεργασία φυσικής γλώσσας.

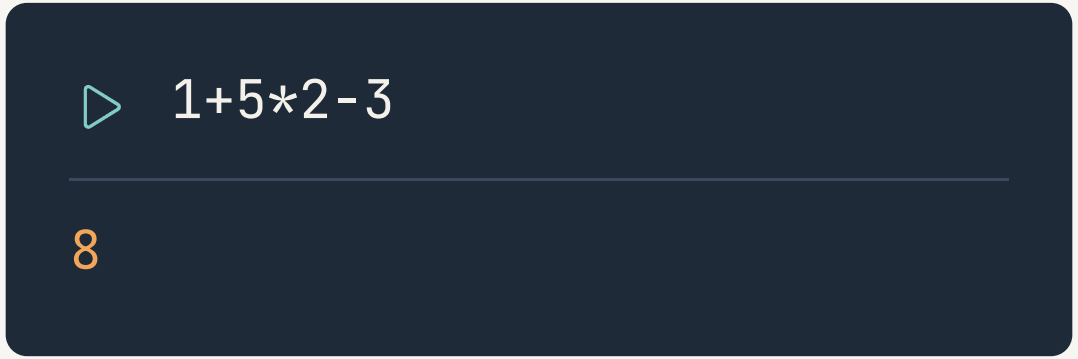
Ξεκίνησε το 2001 στο University of Pennsylvania (Steven Bird, Edward Loper). Συνοδεύεται από δωρεάν βιβλίο: nltk.org/book

Τι περιέχει

- Σώματα κειμένων (corpora)
- Tokenization
- Επισήμανση μερών του λόγου
- Συντακτική ανάλυση
- Συχνότητες και στατιστικά

Δουλεύουμε στο Google Colab

- **File → Save a copy in Drive:** δουλεύετε στο δικό σας αντίγραφο
- **Shift+Enter:** εκτελεί το τρέχον κελί
- **Σειρά:** από πάνω προς τα κάτω
- **Αποσύνδεση;** Runtime → Run all



Κώδικας επάνω, αποτέλεσμα κάτω

Τι κατεβάζουμε στο πρώτο κελί

```
nltk.download(['gutenberg', 'genesis', 'inaugural', 'nps_chat', 'webtext', 'treebank'])
```

Συλλογή	Περιεχόμενο
gutenberg	18 βιβλία του Project Gutenberg: Austen, Melville, Shakespeare...
genesis	Η Γένεση σε 8 εκδοχές: αγγλικά, γαλλικά, γερμανικά, φινλανδικά...
inaugural	60 λόγοι ορκωμοσίας Αμερικανών προέδρων, 1789–2025
nps_chat	10.567 μηνύματα από online chat rooms
webtext	Φόρουμ Firefox, αγγελίες γνωριμιών, σενάρια ταινιών, κριτικές κρασιών
treebank	Απόσπασμα του Penn Treebank: άρθρα της Wall Street Journal

Περίπου 7,5 MB συνολικά · όλα στα αγγλικά

Τα 9 κείμενα του nltk.book

`text1`
Moby Dick
Herman Melville, 1851

`text2`
Sense and Sensibility
Jane Austen, 1811

`text3`
The Book of Genesis
Βίβλος, μετάφραση King James

`text4`
Inaugural Address Corpus
λόγοι ορκωμοσίας Αμερικανών προέδρων

`text5`
Chat Corpus
μηνύματα από chat rooms

`text6`
Monty Python and the Holy Grail
σενάριο κωμωδίας, 1975

`text7`
Wall Street Journal
οικονομικά άρθρα εφημερίδας

`text8`
Personals Corpus
αγγελίες γνωριμιών

`text9`
The Man Who Was Thursday
G. K. Chesterton, 1908

Δύο τρόποι να μετρήσουμε

4 tokens
κάθε εμφάνιση

`very` `very` `very` `nice`

2 types
κάθε διαφορετική λέξη

`very` `nice`

Προσοχή: και η στίξη μετράει. Το «nice.» σπάει σε nice και .

Οι εντολές που θα τρέξουμε

Εντολή	Τι επιστρέφει
len(text2)	πλήθος tokens
set(text2)	σύνολο διαφορετικών λέξεων (types)
sorted(set(text2))	ίδιο σύνολο, αλφαβητικά
len(set(text3))	μέγεθος λεξιλογίου
text2.count('without')	συχνότητα μίας λέξης
len(set(text3)) / len(text3)	λεξιλογική ποικιλία

ΛΕΞΙΛΟΓΙΚΗ ΠΟΙΚΙΛΙΑ · TYPE-TOKEN RATIO

types ÷ tokens

`len(set(text)) / len(text)`

6,2%

στη Γένεση (text3)

Όσο μεγαλύτερη η τιμή, τόσο πλουσιότερο το λεξιλόγιο. Εδώ κάθε λέξη επανέρχεται κατά μέσο όρο 16 φορές.

Προσοχή: η τιμή εξαρτάται από το μήκος

Κείμενο	Tokens	Ολόκληρο	Πρώτα 4.800
Moby Dick (text1)	260.819	0,07	0,34
Αγγελίες γνωριμιών (text8)	4.867	0,23	0,23
Γένεση (text3)	44.764	0,06	0,14

Ολόκληρο, το Moby Dick φαίνεται φτωχότερο από τις αγγελίες· σε ίσο δείγμα περνάει μπροστά.

Συγκρίνουμε πάντα ίσα δείγματα, π.χ. στην εργασία της χρονιάς για ανίχνευση γραφής από AI.

Τι είναι μια κανονική έκφραση

Ένα πρότυπο που περιγράφει ποια κομμάτια ψάχνουμε μέσα σε ένα κείμενο.

Λέξεις

`\w+`

→ Corfu, city, is

Αριθμοί

`\d+`

→ 2026

Ημερομηνίες

`\d+/\d+/\d+`

→ 21/10/2026

Email

`\w+@\w+\.\w+`

→ maria@uni.gr

Η re_show() δείχνει τι ταιριάζει

```
s = "Corfu city is very very very nice. Google is very popular."
nltk.re_show('o+', s)
```

```
C{o}rfu city is very very very nice. G{oo}gle is very p{o}pular.
```

Θα δοκιμάσουμε: `+` `( )` `\s` `|` `[ ]`

Πριν ξεκινήσετε

Τρέχετε τα κελιά με τη σειρά

Απλά εισαγωγικά ' και " — όχι τα ‘ ’ “ ” του Word

Πεζά ≠ κεφαλαία: 'The' ≠ 'the'

r'...' όταν το πρότυπο περιέχει \

Ανοίξτε το notebook

[σύνδεσμος Colab]

Πρώτο βήμα: File → Save a copy in Drive