

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Τμήμα Πληροφορικής



Εργασία Μαθήματος **Επεξεργασία Φυσικής Γλώσσας**

Αριθμός εργασίας – Τίτλος εργασίας	Απαλλακτική εργασία Ιουνίου-Σεπτεμβρίου 2022
Όνομα φοιτητή	Καλλίγερος Αναστάσης
Αρ. Μητρώου	Π19253
Ημερομηνία παράδοσης	18/7/2022



Εκφώνηση εργασίας

B.2 Λύση σε άλλη γλώσσα προγραμματισμού

Εναλλακτικά μπορείτε να χρησιμοποιήσετε άλλες γλώσσες προγραμματισμού ή libraries των κώδικα και την λειτουργικότητα των οποίων πρέπει να τεκμηριώσετε πειστικά και κατανοητά σύμφωνα με τα παρακάτω :

Θέμα 1^ο (20 μονάδες)

Ανάπτυξη Λεκτικού Αναλυτή σε άλλη γλώσσα Προγραμματισμού. Αναζητείστε στο διαδίκτυο ή αναπτύξτε εσείς Λεκτικό Αναλυτή που διαβάζει μια μικρή ιστορία και μπορεί να παράγει μια λίστα από προτάσεις, κάθε μια από τις οποίες περιέχει μια λίστα από λέξεις. Τεκμηριώστε πειστικά τον κώδικά σας.

Θέμα 2^ο (20 μονάδες)

Ανάπτυξη Συντακτικού Αναλυτή σε άλλη γλώσσα Προγραμματισμού. Αναζητείστε στο διαδίκτυο ή αναπτύξτε εσείς Συντακτικό Αναλυτή που με βάση τους κανόνες συντακτικής ανάλυσης της πρότυπης λύσης σε Prolog που σας δόθηκε παράγει το συντακτικό δένδρο της πρότασης. Τεκμηριώστε πειστικά τον κώδικά σας.

Θέμα 3^ο (30 μονάδες)

Ανάπτυξη Σημασιολογικού Αναλυτή σε άλλη γλώσσα Προγραμματισμού. Αναζητείστε στο διαδίκτυο ή αναπτύξτε εσείς Σημασιολογικό Αναλυτή που με βάση τους κανόνες σημασιολογικής ανάλυσης της πρότυπης λύσης σε Prolog που σας δόθηκε παράγει τα σημειώματα της πρότασης (σχέσεις μεταξύ ρημάτων, ουσιαστικών, επιθέτων, κ.λπ). Τεκμηριώστε πειστικά τον κώδικά σας.



Θέμα 4^ο (30 μονάδες)

Πρόγραμμα στην γλώσσα προγραμματισμού της επιλογής σας, για την ενημέρωση και πραγματοποίηση ερωταποκρίσεων σε Βάση Γνώσης που έχετε αναπτύξει για αυτό τον σκοπό. Οι ερωτήσεις και απαντήσεις θα δίδονται σε φυσική γλώσσα.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

1	Εισαγωγή	4
2	Περιγραφή του προγράμματος	4
3	Επίδειξη της λύσης	6
4	Βιβλιογραφικές Πηγές	13



1 Εισαγωγή

Αρχικά, μετά από προσεκτική μελέτη της εκφώνησης αποφάσισα να υλοποιήσω την εργασία επιλέγοντας το B2. Δηλαδή να λύσω τα 4 θέματα που περιλαμβάνονται στο B μέρος της εκφώνησης χρησιμοποιώντας άλλη γλώσσα προγραμματισμού και συγκεκριμένα την python. Επίσης ήταν αναγκαίο για την λειτουργικότητα του κάθε κώδικα να γίνουν install κάποιες libraries όπως η nltk. Ακόμη για την δημιουργία Βάσης Γνώσης χρησιμοποιήθηκε η SQLite όσον αφορά το 4^ο θέμα. Τέλος ως εργαλείο ανάπτυξης του κάθε κώδικα επέλεξα την πιο νέα και πληρέστερη έκδοση του Visual Studio Code. Παρακάτω γίνεται αναλυτική παρουσίαση των θεμάτων με χαρακτηριστικά screenshots για την επίδειξη των λύσεων.

2 Περιγραφή του προγράμματος

Θέμα 1^ο (20 μονάδες)

Για το πρώτο θέμα έπρεπε να φτιάξουμε λεκτικό αναλυτή ώστε διαβάζοντας μια μικρή ιστορία να μπορεί να παραχθεί μια λίστα από προτάσεις και στη συνέχεια μια λίστα από λέξεις που περιέχουν αυτές οι προτάσεις. Αρχικά κάναμε import το nltk που αποτελεί μια εργαλειοθήκη για την επεξεργασία φυσικής γλώσσας (natural language toolbox kit). Από αυτή την εργαλειοθήκη κάναμε import τα sent_tokenize και word_tokenize που χωρίζουν το αρχείο randomText.txt σε προτάσεις και λέξεις αντίστοιχα. Έτσι κάνουμε print τις προτάσεις και στη συνέχεια τις λέξεις κάθε πρότασης του αρχείου randomText.txt.

Θέμα 2^ο (20 μονάδες)

Στο δεύτερο θέμα καλούμαστε να δημιουργήσουμε έναν συντακτικό αναλυτή που θα παράγει συντακτικό δέντρο. Όπως και στο πρώτο ερώτημα κάνουμε import το πακέτο nltk καθώς και τα sent_tokenize, word_tokenize και PunktSentenceTokenizer που ουσιαστικά διαιρεί το .txt αρχείο σε μια λίστα από προτάσεις. Επίσης συμπεριέλαβα σαν σχόλιο μέσα στον κώδικα την λίστα με τα tags που αντιστοιχούν σε κάθε μέρος του λόγου. Στη συνέχεια με την συνάρτηση process_content κάνω print κάθε λέξη μέσα στο randomText.txt μαζί με την συντακτική σημασία κάθε λέξης. Τέλος με την γραμμή chunked.draw() κάνουμε print το συντακτικό δέντρο προσθέτοντας κάθε πρόταση.



Θέμα 3^ο (30 μονάδες)

Για τον σημασιολογικό αναλυτή χρησιμοποίησα 2 .txt αρχεία τα οποία περιείχαν θετικές και αρνητικές κριτικές ταινιών(negative.txt, positive.txt). Στη συνέχεια όρισα να λαμβάνω μόνο τα μέρη του λόγου που είναι ρήματα, επίθετα και επιρρήματα για να περιορίσω λίγο τα δεδομένα και να λειτουργήσουν αποδοτικότερα οι αλγόριθμοι. Για την χρήση των αλγορίθμων έκανα import το sklearn που είναι ένα package που περιέχει διάφορους αλγόριθμους και έκανα train και test τα διάφορα δεδομένα. Έπειτα εμφανίζω τις λέξεις (ρήμα, επίθετο ή επίρρημα) που εμφανίζονται περισσότερες φορές στα θετικά σε σχέση με τα αρνητικά και το αντίστροφο.

Θέμα 4^ο (30 μονάδες)

Το πρώτο πράγμα που έκανα είναι να κατασκευάσω μια βάση γνώσης. Η βάση γνώσης είναι στην ουσία μια Βάση δεδομένων (Database) που περιέχει πίνακες στους οποίους είναι καταχωρημένη όλη η σχετική γνώση. Για το παράδειγμα μας λοιπόν κατασκευάζουμε μια βάση (σε sqlite) που την ονομάζουμε Knowledge και καταχωρούμε την γνώση σε ένα πίνακα της βάσης που το ονομάζουμε knowledge.



3 Επίδειξη της λύσης

. Θέμα 1^ο (20 μονάδες)

Δημιουργία Λεκτικού Αναλυτή:

thema1.py

```
thema1.py X
C: > Users \tasso > Desktop > EPEXERGASIAFYSIKHSGUWSSAS > tema1.py
1  #-*- coding: utf-8 -*-
2  """
3  Created on Fri Jul 15 15:46:46 2022
4
5  @author: Kalligeros Anastasis
6  """
7  #arxika kanw import to natural language toolkit
8  import nltk
9
10
11
12  # import sentence tokenizer kai word tokenizer
13  from nltk.tokenize import sent_tokenize, word_tokenize
14
15  #----- 10 Thema -----#
16
```

Try the new cross-platform PowerShell <https://aka.ms/powershell>

```
PS C:\Users\tasso> python -u "c:\Users\tasso\Desktop\EPEXERGASIAFYSIKHSGUWSSAS\thema1.py"
```

List of sentences:

```
['Lorem Ipsum is simply dummy text of the printing and typesetting industry.', 'Lorem Ipsum has been the industry standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book.', 'It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged.']
```

List of words and punctuations:

```
['Lorem', 'Ipsum', 'is', 'simply', 'dummy', 'text', 'of', 'the', 'printing', 'and', 'typesetting', 'industry', '.', 'Lorem', 'Ipsum', 'has', 'been', 'the', 'industry', 'standard', 'dummy', 'text', 'ever', 'since', 'the', '1500s', ',', 'when', 'an', 'unknown', 'printer', 'took', 'a', 'galley', 'of', 'type', 'and', 'scrambled', 'it', 'to', 'make', 'a', 'type', 'specimen', 'book', '.', 'It', 'has', 'survived', 'not', 'only', 'five', 'centuries', ',', 'but', 'also', 'the', 'leap', 'into', 'electronic', 'typesetting', ',', 'remaining', 'essentially', 'unchanged', '.']
```

PS C:\Users\tasso>



Θέμα 2^ο (20 μονάδες)

Δημιουργία Συντακτικού Αναλυτή:

thema2.py

```
File Edit Selection View Go Run Terminal Help
thema2.py - Visual Studio Code

thema2.py X
C:\Users\tasso\Desktop\EPEXERGASIAFYSIKHSGLWSSAS> tema2.py > ...
1  # -*- coding: utf-8 -*-
2  """
3  Created on Fri Jul 15 19:33:18 2022
4
5  @author: Kalligeros Anastasis
6  """
7  #arxika kanw import to natural language toolkit
8  import nltk
9  #kanw import to PunktSentenceTokenizer poy diairei to.txt arxeio
10 #se mia lista apo sentences me xrhsh algorithmou
11 from nltk.tokenize import PunktSentenceTokenizer, sent_tokenize, word_tokenize
12 from nltk.corpus import state_union
13
14 #----- 2o Thema -----#
15
16

TERMINAL JUPYTER PROBLEMS OUTPUT DEBUG CONSOLE
Code + - + ^ x

Windows PowerShell
Copyright (c) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS C:\Users\tasso> python -u "C:\Users\tasso\Desktop\EPEXERGASIAFYSIKHSGLWSSAS\thema2.py"
[('Lorem', 'NNP'), ('Ipsum', 'NNP'), ('is', 'VBZ'), ('simply', 'RB'), ('dummy', 'JJ'), ('text', 'NN'), ('of', 'IN'), ('the', 'DT'), ('printing', 'NN'), ('and', 'CC'), ('typesetting', 'NN'), ('industry', 'NN'), ('.', '.')]

Ln 18, Col 62 Spaces: 4 UTF-8 CRLF Python 3.10.4 64-bit
```



thema2.py - Visual Studio Code

```
1 # -*- coding: utf-8 -*-
2
3 Created on Fri Jul 15 19:33:18 2022
4
5 @author: Kalligeros Anastasis
6
7 #arxika kanw import to natural language toolkit
8 import nltk
9 #kanw import to PunktSentenceTokenizer poy diairei to.txt arxelo
10 #se mia lista apo sentences me xrhsh algorithmou
11 from nltk.tokenize import PunktSentenceTokenizer, sent_tokenize, word_tokenize
12 from nltk.corpus import state_union
13
14 #----- 2o Thema -----#
15
16
```

TERMINAL JUPYTER PROBLEMS OUTPUT DEBUG CONSOLE

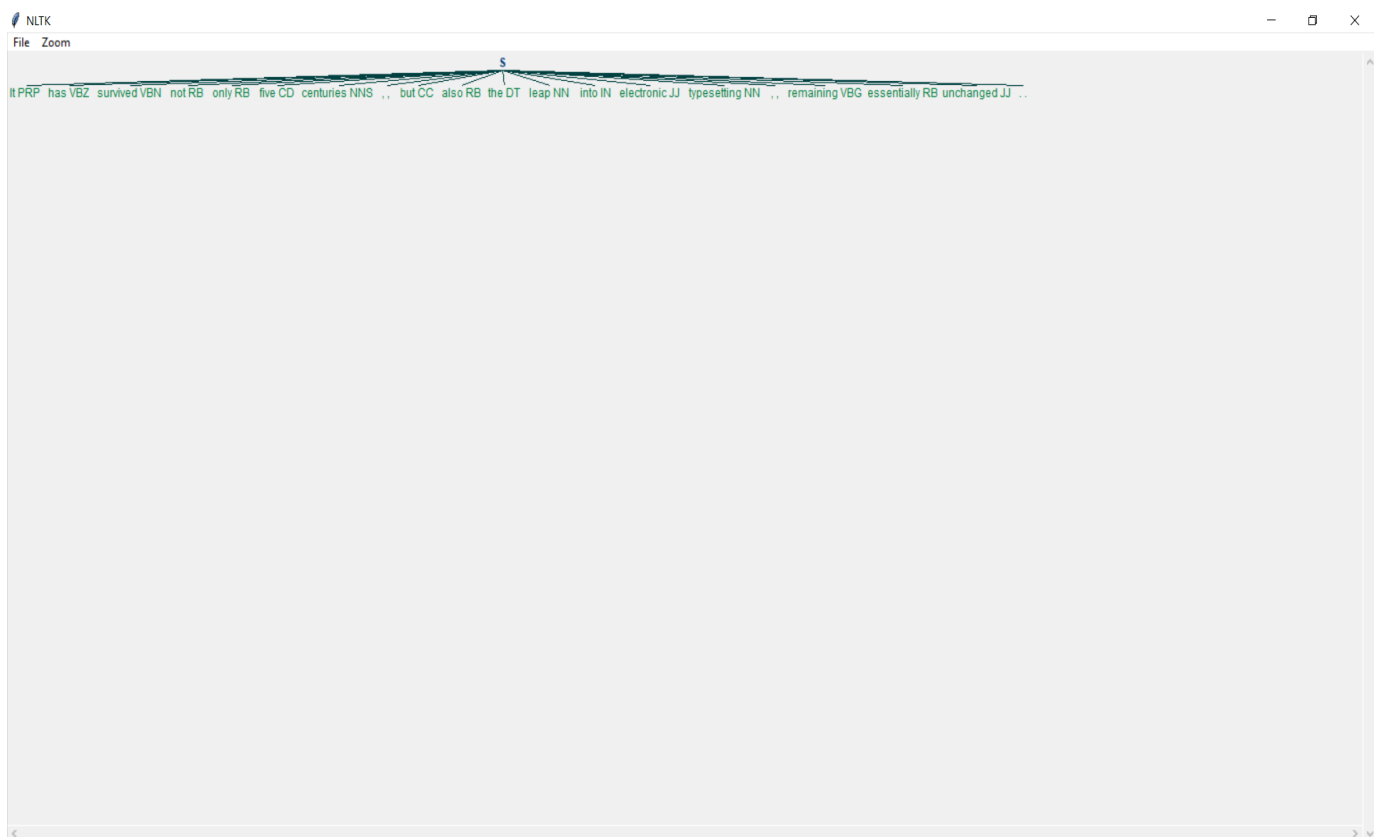
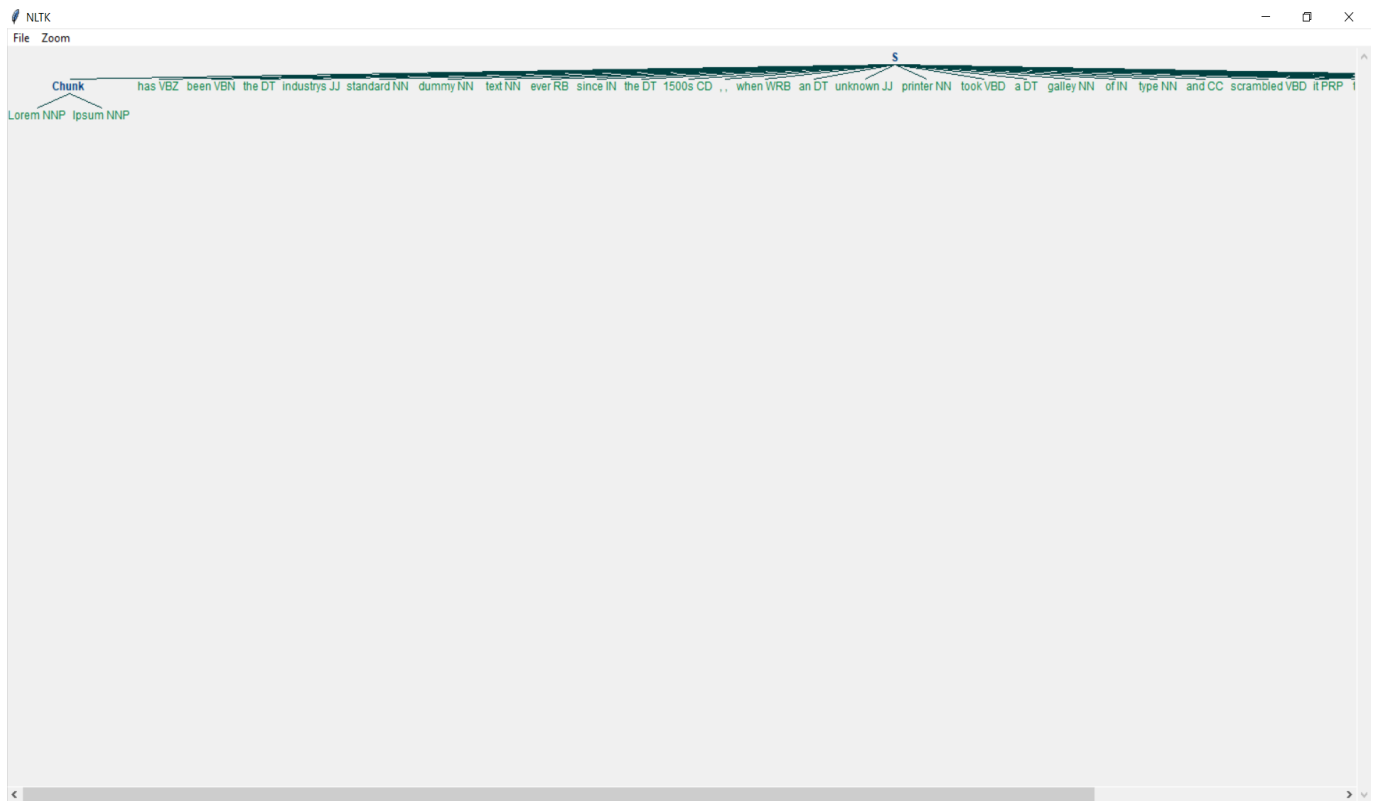
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell <https://aka.ms/pscore6>

```
PS C:\Users\tasso> python -u "C:\Users\tasso\Desktop\EPEXERGASIAFYSIKHSGLWSSAS\thema2.py"
[('Lorem', 'NNP'), ('Ipsum', 'NNP'), ('is', 'VBZ'), ('simply', 'RB'), ('dummy', 'JJ'), ('text', 'NN'), ('of', 'IN'), ('the', 'DT'), ('printing', 'NN'), ('and', 'CC'), ('typesetting', 'NN'), ('industry', 'NN'), ('.', '.')]
[('Lorem', 'NNP'), ('Ipsum', 'NNP'), ('has', 'VBZ'), ('been', 'VBN'), ('the', 'DT'), ('industrys', 'JJ'), ('standard', 'NN'), ('dummy', 'NN'), ('text', 'NN'), ('ever', 'RB'), ('since', 'IN'), ('the', 'DT'), ('1500s', 'CD'), ('.', '.'), ('when', 'WRB'), ('an', 'DT'), ('unknown', 'JJ'), ('printer', 'NN'), ('took', 'VBD'), ('a', 'DT'), ('galley', 'NN'), ('of', 'IN'), ('type', 'NN'), ('and', 'CC'), ('scrambled', 'VBD'), ('it', 'PRP'), ('to', 'TO'), ('make', 'VB'), ('a', 'DT'), ('type', 'NN'), ('specimen', 'NNS'), ('book', 'NN'), ('.', '.')]

```

Ln 18, Col 62 Spaces: 4 UTF-8 CRLF Python 3.10.4 64-bit





```
File Edit Selection View Go Run Terminal Help
thema2.py - Visual Studio Code

thema2.py X
C:\Users\tasso\Desktop\EPEXERGASIAFYSIKHSGLWSSAS > tema2.py > ...
1  # -*- coding: utf-8 -*-
2  """
3  Created on Fri Jul 15 19:33:18 2022
4
5  @author: Kalligeros Anastasis
6  """
7  #arxika kanw import to natural language toolkit
8  import nltk
9  #kanw import to PunktSentenceTokenizer poy diairei to.txt arxeio
10 #se mia lista apo sentences me xrhsh algorithmou
11 from nltk.tokenize import PunktSentenceTokenizer, sent_tokenize, word_tokenize
12 from nltk.corpus import state_union
13
14 #----- 2o Thema -----#
15
16
17 #anoigw to arxeio .txt kai to diavazw
18 file = open("C:/Users/tasso/Desktop/EPEXERGASIAFYSIKHSGLWSSAS/randomText.txt","r")
19 #an to arxeio einai se read mode
20 if file.mode == 'r':
...

TERMINAL JUPYTER PROBLEMS OUTPUT DEBUG CONSOLE
Windows PowerShell
Copyright (c) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS C:\Users\tasso> python -u "C:\Users\tasso\Desktop\EPEXERGASIAFYSIKHSGLWSSAS\thema2.py"
[('Lorem', 'NMP'), ('Ipsum', 'NMP'), ('is', 'VBZ'), ('simply', 'RB'), ('dummy', 'JJ'), ('text', 'NN'), ('of', 'IN'), ('the', 'DT'), ('printing', 'NN'), ('and', 'CC'), ('typesetting', 'NN'), ('industry', 'NN'), ('.', '.')]
[('Lorem', 'NMP'), ('Ipsum', 'NMP'), ('has', 'VBZ'), ('been', 'VBN'), ('the', 'DT'), ('industries', 'JJ'), ('standard', 'NN'), ('dummy', 'NN'), ('text', 'NN'), ('ever', 'RB'), ('since', 'IN'), ('the', 'DT'), ('1500s', 'CD'), ('.', '.'), ('when', 'WRB'), ('an', 'DT'), ('unknown', 'JJ'), ('printer', 'NN'), ('took', 'VBD'), ('a', 'DT'), ('galley', 'NN'), ('of', 'IN'), ('type', 'NN'), ('and', 'CC'), ('scrambled', 'VBD'), ('it', 'PRP'), ('to', 'TO'), ('make', 'VB'), ('a', 'DT'), ('type', 'NN'), ('specimen', 'NNS'), ('book', 'NN'), ('.', '.')]
[('It', 'PRP'), ('has', 'VBZ'), ('survived', 'VBN'), ('not', 'RB'), ('only', 'RB'), ('five', 'CD'), ('centuries', 'NNS'), ('.', '.'), ('but', 'CC'), ('also', 'RB'), ('the', 'DT'), ('leap', 'NN'), ('into', 'IN'), ('electronic', 'JJ'), ('typesetting', 'NN'), ('.', '.'), ('remaining', 'VBG'), ('essentially', 'RB'), ('unchanged', 'JJ'), ('.', '.')]
PS C:\Users\tasso>
```



Θέμα 3^ο (30 μονάδες)

Δημιουργία Σηματολογικού Αναλυτή:

naturalLang.py

```
187
188
189 save_classifier = open("LinearSVC.pickle","wb")
190 pickle.dump(LinearSVC_classifier, save_classifier)
191 save_classifier.close()
192
193
194 NuSVC_classifier = SklearnClassifier(NuSVC())
195 NuSVC_classifier.train(training_set)
```

Most Informative Features

boring = True	neg : pos	=	19.8 : 1.0
dull = True	neg : pos	=	15.0 : 1.0
engrossing = True	pos : neg	=	13.4 : 1.0
worse = True	neg : pos	=	12.0 : 1.0
captures = True	pos : neg	=	12.0 : 1.0
wonderful = True	pos : neg	=	12.0 : 1.0
thoughtful = True	pos : neg	=	10.6 : 1.0
delightful = True	pos : neg	=	9.9 : 1.0
flawed = True	pos : neg	=	9.9 : 1.0
thin = True	neg : pos	=	8.8 : 1.0
absorbing = True	pos : neg	=	8.6 : 1.0
inventive = True	pos : neg	=	8.6 : 1.0
refreshing = True	pos : neg	=	8.6 : 1.0
contrived = True	neg : pos	=	8.1 : 1.0
mildly = True	neg : pos	=	8.1 : 1.0

Naive Bayes Algo accuracy percent: 70.46257062146893
MNB_classifier accuracy percent: 69.8093220338983
BernoulliNB_classifier accuracy percent: 70.10946327683615
LogisticRegression_classifier accuracy percent: 69.35028248587571
SGDClassifier_classifier accuracy percent: 68.07909604519774
LinearSVC_classifier accuracy percent: 67.70833333333334
NuSVC_classifier accuracy percent: 72.17514124293785
PS C:\Users\tasso

Όπως βλέπετε η λέξη dull θα εμφανιστεί 15 φορές στις αρνητικές κριτικές προς 1 που θα εμφανιστεί στις θετικές το οποίο είναι λογικό. Πιο κάτω βλέπουμε τους διάφορους αλγόριθμους και τις επί τοις 100 αποδόσεις τους. Πιο αποδοτικός στη συγκεκριμένη εκτέλεση είναι ο αλγόριθμος NuSVC(Nu-Support Vector Classification) με 72.17% απόδοση. Τέλος χρησιμοποιούμε κάνοντας import το pickle για να αποθηκεύουμε τους αλγορίθμους μας και να είναι πιο αποδοτικό και γρήγορο το πρόγραμμά μας.



Θέμα 4^ο (30 μονάδες)

Ο πίνακας που κατασκευάστηκε είναι ο εξής:

DB Browser for SQLite - C:\Users\tasso\Desktop\EPEXERGASIAFYSIKHSGLWSSAS\Knowledge.db

File Edit View Tools Help

New Database Open Database Write Changes Revert Changes Open Project Save Project

Database Structure Browse Data Edit Pragmas Execute SQL

Table: knowledge

	id	Name	Verb	Noun	Adjective	Intransitive_Verb	Transitive_Verb	Adverb	Receiver
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	1	anna	NULL	dog	NULL	run	give	slowly	john
2	2	kostas	NULL	book	NULL	run	give	NULL	tomy
3	3	dimitris	speak	phone	NULL	run	NULL	quickly	NULL
4	4	nikos	listens	sounds	tall	run	NULL	quickly	kostas
5	5	katerina	watch	tv	slim	run	give	slowly	nikoleta
6	6	nikoleta	play	football	blonde	run	NULL	NULL	nikolea
7	7	maria	love	book	short	run	NULL	slowly	NULL
8	8	giannis	need	food	black	run	give	NULL	dimitris
9	9	thodoris	have	food	fat	run	NULL	quickly	konstantina
10	10	apostolis	hate	spiders	NULL	NULL	give	NULL	thodoris
11	11	matina	chase	cat	NULL	NULL	give	NULL	NULL
12	12	konstantina	NULL	NULL	scary	run	NULL	quickly	matina

Στο αρχείο gr_voc.fcfg έχουμε δημιουργήσει μια γραμματική και ένα λεξιλόγιο. Η γραμματική μας όπως έχει σχεδιαστεί δημιουργεί ερωτήματα (Queries) για την βάση γνώσης (Knowledge Base) τα οποία αφορούν την απάντηση του εκάστοτε ερωτήματος που τίθεται από τον χρήστη.



QA.py

```
File Edit Selection View Go Run Terminal Help
QA.py - Visual Studio Code

C:\Users\tasso\Desktop\EPEXERGASIAFYSIKHSGLWSSAS > QA.py
4 from nltk import load_parser
5 cp = load_parser('gr_voc.fcfg')
6 query = 'who loves book'
7 trees = list(cp.parse(query.split()))
8 answer = trees[0].label()['SEM']
9 answer = [s for s in answer if s]
10 q = ' '.join(answer)
11 print(q)
12 sqliteConnection = sqlite3.connect('C:/Users/tasso/Desktop/EPEXERGASIAFYSIKHSGLWSSAS/Knowledge.db')
13 cursor = sqliteConnection.cursor()
14 cursor.execute(q+';')
15 records = cursor.fetchall()
16 for row in records:
17     print(row[0])
18 sqliteConnection.close()

TERMINAL JUPYTER PROBLEMS OUTPUT DEBUG CONSOLE
Python Debug Console + - [ ] [x] [y] [z]

Windows PowerShell
Copyright (c) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS C:\Users\tasso\Desktop\EPEXERGASIAFYSIKHSGLWSSAS> & 'C:\Python310\python.exe' 'c:\Users\tasso\.vscode\extensions\ms-python.python-2022.10.1\pythonFiles\lib\python\debugpy\adapter\..\debugpy\launcher' '50618' '-.' 'c:\Users\tasso\Desktop\EPEXERGASIAFYSIKHSGLWSSAS\QA.py'
SELECT Name FROM Knowledge WHERE Verb='love' AND Noun='book'
mary
PS C:\Users\tasso\Desktop\EPEXERGASIAFYSIKHSGLWSSAS>
```

Εκτελούμε το πρόγραμμα μας για να δούμε την έξοδο που προκύπτει σύμφωνα με το ερώτημα που υποβάλλαμε.

4 Βιβλιογραφικές Πηγές

Σημειώσεις GUNET2 (ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ).