Γλωσσική Τεχνολογία

Εργαστηριακή Άσκηση Σεπτέμβριος 2010

Γιαννουδάκης Ιωάννης AM:3640 Έτος:5ο

Παπαρροδοπούλου Αναστασία AM:3873 Έτος:5ο

Χαντζής Φώτης AM:3771 Έτος:5ο

ΜΕΡΟΣ Α'

Σκοπός του πρώτου μέρους ήταν η δημιουργία ενός ανεστραμμένου ευρετηρίου για μία συλλογή 1000 κειμένων από την Wikipedia. Στο ευρετήριο αυτό, που έχει XML μορφή, υποβάλλαμε μία σειρά από ερωτήματα με σκοπό να τα αναζητήσουμε και να μετρήσουμε το μέσο χρόνο αναζήτησής μέσα σε αυτό. Για την υλοποίηση του ευρετηρίου εκτελέσαμε τα παρακάτω βήματα.

Tokenization

Αναζητήσαμε τα tokens και τα σημεία στίξης σε κάθε ένα από τα κείμενα της συλλογής, ενώ αποκλείσαμε νέες γραμμές, tabs και whitespaces. Η διαδικασία έγινε διαβάζοντας το κάθε αρχείο γραμμή προς γραμμή. Για το tokenization της κάθε γραμμής διατρέχαμε τους χαρακτήρες κάθε γραμμής και όσο δεν συναντούσαμε κάποιο σημείο στίξης ή κάποιο χαρακτήρα από το σύνολο {\n,\r,\t,' '} προσθέταμε τον εκάστοτε χαρακτήρα σε μία μεταβλητή, token, που χρησιμοποιούμε. Όταν συναντούσαμε ένα από τους παρακάτω χαρακτήρες:

```
delimiters = ('.', '\'', '`', '<', '>', '{', '}', '[', ']', '(', ')', '\\',\
        '?',  ';', ':', '~', '"', '!', '%', '#', '@', '$', '^', '&', \
        '-',  '*', '+', '|', ',', '...', '/')
```

κάναμε append στην λίστα των tokens τόσο το χαρακτήρα αυτό καθώς και το μέχρι τώρα αναγνωρισμένο token. Έπειτα η μεταβλητή token αρχικοποιείται σε κενό ώστε να αναγνωρίσουμε το νέο επόμενο token. Τα tokens του κάθε αρχείου αποθηκεύονται σε ένα νέο αρχείο με όνομα ίδιο με το αρχικό συν το πρόθεμα out_ , στον φάκελο out_tokenized. Το tokenization γίνεται τρέχοντας tokenize.py από την γραμμή εντολών.

Part Of Speech Tagging

Για τον μορφοσυντακτικό σχολιασμό των tokens χρησιμοποιήσαμε τον TreeTagger, καθώς η υλοποίηση έγινε σε Windows 7. Η εκτέλεση του TreeTagger γίνεται γράφοντας:

```
# command for tree-tagger
cmd = "tree-tagger.exe -quiet -token -lemma english.par "
```

Συμπληρώνοντας την εντολή με τα path names των αρχείων εισόδου και εξόδου εκτελούμε το POStagging μέσω του module os.

```
# execute tree-tagger for the specified file
os.system(cmd)
```

Η εκτέλεση του Tree Tagger γίνεται στο αρχείο postagging.py. Το αποτέλεσμα για κάθε tokenized κείμενο αποθηκεύεται σε ένα νέο αρχείο με το πρόθεμα Tagged_ στο φάκελο tagged_texts.

Αναπαράσταση στο Διανυσματικό Χώρο

Με βάση τους πίνακες των open και closed PoSTags, αναπαραστήσαμε διανυσματικά τα κείμενα που προκύπτουν από την μορφοσυντακτική ανάλυση. Για κάθε tagged-κείμενο, αφαιρέσαμε τους τερματικούς όρους και εκείνα τα λήμματα/tokens που η γραμματική τους κατηγορία είναι μία από τις παρακάτω.

```
closed_class = ('CD', 'CC', 'DT', 'EX', 'IN', 'LS', 'MD', 'PDT', 'POS', 'PRP',\
        'PRP$', 'RP', 'TO', 'UH', 'WDT', 'WP', 'WP$', 'WRB', 'SYM', '\'\'', '``')
```

Επίσης αγνοήθηκαν και tokens που σαν λήμμα είχαν ένα από τα παρακάτω:

```
ingored_lemmas = ('%', '$', '%', '<unknown>', '—','.')
```

Αφού έγινε η παραπάνω διαλογή των λημμάτων, έγινε η μετατροπή τους σε πεζά, και έπειτα από ταξινόμησή τους με βάση το λήμμα, υπολογίσαμε για κάθε ένα την συχνότητα εμφάνισης του στο κείμενο. Η συνάρτηση που αναλαμβάνει την αφαίρεση των stop words, των unknown λημμάτων και των σημείων στίξης, είναι η remove_stopWords στο αρχείο tokenizing.py.

Η δημιουργία της διανυσματικής μορφής έχει ως εξής:

1. Αρχικοποιούμε μία λίστα με τα vectors (vec_list)
2. Γίνεται ταξινόμηση με βάση το λήμμα. Έτσι τα λήμματα ομαδοποιούνται ώστε tokens με ίδιο λήμμα να είναι συνεχόμενα.
3. Διατηρείται μία βοηθητική λίστα prev με το token που ελέγχτηκε τελευταίο.
4. Για κάθε τριάδα (token tag lemma) της open λίστας:
   - *Αν* το λήμμα της τρέχον τριάδας είναι ίδιο με το λήμμα της prev
     - Αυξάνουμε την συχνότητα του λήμματος που εισάγαμε τελευταίο κατά 1
   - *Αλλιώς*
     - Δημιουργούμε ένα dictionary, με πεδία ‘lemma’, ‘pos_tag’, ‘freq’
     - Αναθέτουμε τιμές σε αυτά
     - Κάνουμε append στην vec_list το νέο dictionary

Τα vectors που προκύπτουν για κάθε αρχείο αποθηκεύονται σε ξεχωριστά αρχεία με όνομα ίδιο με το όνομα του αντίστοιχου tagged κειμένου συν το πρόθεμα ‘vectored_’ , στον φάκελο vector_texts. Για τη δημιουργία της διανυσματικής αναπαράστασης για όλα τα κείμενα της συλλογής, εκτελούμε το script vectorize.py .

Δημιουργία Ανεστραμμένου Ευρετηρίου και ανάθεση βαρών

Τα βήματα 4 και 5 υλοποιήθηκαν μαζί, στο αρχείο indexing.py.

Αρχικά γίνεται η δημιουργία ενός ανεστραμμένου ευρετηρίου κρατώντας σε μία δομή όλα τα λήμματα της συλλογής μαζί με τα id's των κειμένων που εμφανίζονται και την συχνότητα εμφάνισης στο κάθε κείμενο. Σαν δομή επιλέξαμε ένα dictionary. Η υλοποίηση της πρώτης αυτής έκδοσης του ευρετηρίου γίνεται στην συνάρτηση inverted_index. Διατρέχουμε κάθε τριάδα (lemma PosTag Frequency) κάθε vector κειμένου. Για κάθε τριάδα κρατάμε σε μία βοηθητική μεταβλητή την τιμή του λήμματος και δημιουργούμε ένα ζευγάρι με : το id του κειμένου στο οποίο βρίσκεται και την συχνότητα του. Επίσης επεκτείνουμε το ζευγάρι με το βάρος του λήμματος, που αρχικοποιείται σε 0.0 . Αν το λήμμα υπάρχει στο dictionary, τότε το ζευγάρι του, γίνεται append στην θέση του dictionary όπου βρίσκεται το λήμμα. Διαφορετικά προσθέτουμε μία νέα καταχώρηση στο dictionary.

Στην συνέχεια, υπολογίσαμε χρησιμοποιώντας τη μετρική TF-IDF το βάρος καθενός από τα λήμματα που έχουμε βάλει στο ευρετήριο. Για τον υπολογισμό της μετρικής αυτής υπολογίσαμε τα παρακάτω:

- *Term Frequency*

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

  Όπου $tf_{i,j}$ είναι η συχνότητα ενός όρου i στο κείμενο j και δίνεται διαιρώντας το πλήθος των εμφανίσεων του όρου με το άθροισμα των εμφανίσεων όλων των όρων του κειμένου.

- *Inverse Document Frequency*

$$idf_j = \log_{10}\left(\frac{\text{αριθμος κειμένων συλλογής}}{\text{αριθμός κειμένων συλλογής που εμφανίζεται ο όρος}}\right)$$

Το βάρος ενός λήμματος προκύπτει ως το γινόμενο των δύο παραπάνω ποσοτήτων. Επίσης κανονικοποιούμε το βάρος του όρου, ώστε μεγάλα κείμενα να μην παίρνουν μεγάλα βάρη. Τέλος, το ευρετήριο γράφεται στο δίσκο σε μορφή XML.

Αξιολόγηση Ευρετηρίου

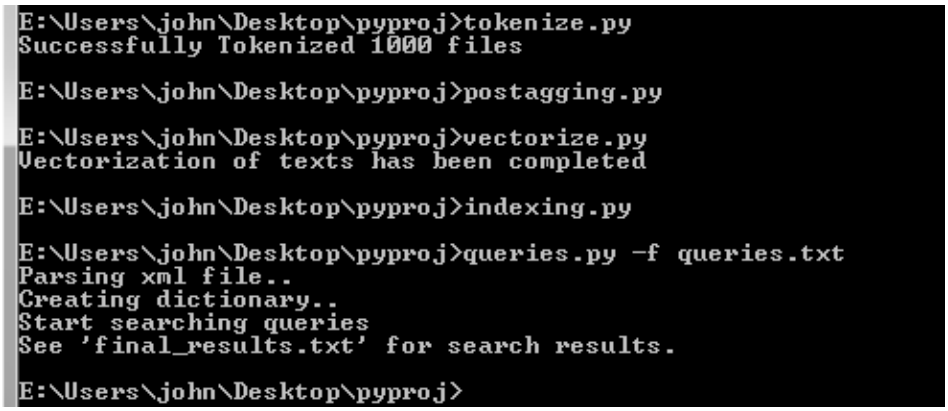
Η αξιολόγηση του ευρετηρίου γίνεται αναζητώντας κάποιες λέξεις μέσα σε αυτό. Οι λέξεις προς αναζήτηση δίνονται στο αρχείο queries.txt. Η αξιολόγηση γίνεται εκτελώντας

queries.py –f queries.txt

Αρχικά γίνεται το parsing του XML ευρετηρίου με τη χρήση του xml.etree.ElementTree module. Έπειτα γίνεται η φόρτωση του xml σε ένα λεξικό ώστε να γίνει η αναζήτηση. Αν κάποιο ερώτημα αποτελείται από μία μόνο λέξη τότε η single_qry_handler επιστρέφει την λίστα με τα docids και τα βάρη στο καθένα. Γίνεται φθίνουσα ταξινόμηση της λίστας με βάση το βάρος και αποθηκεύουμε τα 10 πρώτα αποτελέσματα. Αν το ερώτημα αποτελείται

από περισσότερες από μία λέξεις, τότε καλώντας την single_qry_handler για κάθε λέξη του ερωτήματος, συνδυάζουμε της επιμέρους λίστες ώστε να βρούμε τα κείμενα στα οποία εμφανίζονται όλες οι λέξεις του ερωτήματος. Τα αποτελέσματα της αναζήτησης αποθηκεύονται στο final_results.txt.

ScreenShot από εκτέλεση βημάτων δημιουργίας ευρετηρίου και υποβολής ερωτημάτων σε αυτό:



Απαντήσεις στα ερωτήματα Α' Μέρους

Συνολικά Λήμματα

Τα συνολικά λήμματα της συλλογής υπολογίστηκαν στο αρχείο vectorize.py. Ο αριθμός των συνολικών λημμάτων αποθηκεύεται στην μεταβλητή total_lemmas και γράφεται στο αρχείο 'total_lemmas.txt'. Για τον υπολογισμό του αριθμού συνολικών λημμάτων σε ένα διπλό loop που διατρέχουμε τα vectorized κείμενα, αθροίζουμε τις συχνότητες εμφάνισης κάθε λήμματος κάθε κειμένου (εκτός από τα stop words). Έτσι έχουμε:

Number of total lemmas is: 1684908

Μοναδικά Λήμματα

Ο αριθμός των μοναδικών λημμάτων της συλλογής υπολογίζεται μετρώντας τα lemma tags που υπάρχουν στο index.xml αρχείο και γράφεται στο αρχείο 'unique_lemmas.txt'.

Number of unique lemmas is: 46328

Αποτελέσματα Αναζήτησης

Οι χρόνοι αναζήτησης των queries καθώς και τα αποτελέσματα της αναζήτησης παρουσιάζονται παρακάτω:

Στο αρχείο time_result.txt:

Total Search Time: 0.0329999923706 seconds.

Average Search time: 0.000659999847412 seconds.

Στο αρχείο 'final_results.txt' είναι αποθηκευμένα τα 10 πρώτα αποτελέσματα { docid, weight } για κάθε ερώτημα του queries.txt:

research:

doc_id: 155045 weight: 0.00392278184594
doc_id: 146078 weight: 0.00338711198665
doc_id: 167354 weight: 0.00266640087052
doc_id: 140699 weight: 0.00261432758818
doc_id: 142950 weight: 0.00251548865194
doc_id: 167774 weight: 0.00215899424108
doc_id: 146717 weight: 0.00164217661343
doc_id: 169409 weight: 0.00156852417396
doc_id: 154290 weight: 0.00150278761706
doc_id: 146072 weight: 0.00133193279019

organization:

doc_id: 154826 weight: 0.00254885602848
doc_id: 148131 weight: 0.001642540973
doc_id: 160667 weight: 0.00137787459956
doc_id: 146717 weight: 0.00133394236595
doc_id: 160996 weight: 0.00132054552727
doc_id: 151053 weight: 0.00101929150887
doc_id: 167854 weight: 0.00101816397401
doc_id: 159727 weight: 0.00100580161396
doc_id: 168482 weight: 0.000969738354538
doc_id: 164603 weight: 0.000967505500183

model:

doc_id: 147874 weight: 0.00405270531485
doc_id: 163062 weight: 0.00342962476563
doc_id: 143608 weight: 0.00321358169769
doc_id: 162411 weight: 0.00258121320619
doc_id: 159083 weight: 0.00251387219604
doc_id: 173290 weight: 0.00245802242936
doc_id: 152205 weight: 0.00244654706788
doc_id: 173198 weight: 0.00209565631636
doc_id: 148131 weight: 0.00191523857595
doc_id: 164332 weight: 0.00188764127066

union:

doc_id: 165304 weight: 0.00612499429718
doc_id: 162279 weight: 0.00218807634619
doc_id: 140460 weight: 0.00212315959595
doc_id: 153481 weight: 0.00160301047329
doc_id: 146410 weight: 0.00156811283493
doc_id: 171745 weight: 0.00155652233942
doc_id: 150186 weight: 0.00140728341155
doc_id: 153992 weight: 0.00132888591751
doc_id: 148407 weight: 0.00132797259385
doc_id: 171473 weight: 0.00126225714457

train:

doc_id: 169004 weight: 0.00720865876213
doc_id: 150183 weight: 0.00558013986941

doc_id: 160774 weight: 0.00297979148144
doc_id: 142175 weight: 0.00235716875842
doc_id: 171742 weight: 0.00223495791949
doc_id: 165904 weight: 0.00166748986162
doc_id: 162770 weight: 0.00156447054364
doc_id: 172174 weight: 0.00156207960547
doc_id: 172759 weight: 0.00136959108634
doc_id: 143945 weight: 0.00125021565597

education:
doc_id: 159405 weight: 0.00322119552837
doc_id: 151712 weight: 0.00265377231478
doc_id: 163464 weight: 0.00224689768204
doc_id: 160591 weight: 0.00217685084839
doc_id: 163157 weight: 0.00177896637874
doc_id: 142950 weight: 0.00155819998831
doc_id: 159932 weight: 0.00134391112868
doc_id: 147063 weight: 0.00124277408275
doc_id: 148565 weight: 0.00123763047782
doc_id: 167774 weight: 0.00122274037299

purpose:
doc_id: 148565 weight: 0.00260529023005
doc_id: 168850 weight: 0.00142025349935
doc_id: 170717 weight: 0.000930317992824
doc_id: 139114 weight: 0.000811440644591
doc_id: 172881 weight: 0.000791750514007
doc_id: 154154 weight: 0.000711757350709
doc_id: 166653 weight: 0.000689205839319
doc_id: 146704 weight: 0.000647389987957
doc_id: 167200 weight: 0.000539900415822
doc_id: 153977 weight: 0.000532595062258

language:
doc_id: 153465 weight: 0.014718219387
doc_id: 146609 weight: 0.00836271264516
doc_id: 160538 weight: 0.0070055383384
doc_id: 167941 weight: 0.00677444704876
doc_id: 158234 weight: 0.0066309846486
doc_id: 156920 weight: 0.00523259323886
doc_id: 140914 weight: 0.00466585507006
doc_id: 166314 weight: 0.00464866088079
doc_id: 153895 weight: 0.00431987419557
doc_id: 166929 weight: 0.00394047347987

church:
doc_id: 168130 weight: 0.0163698025899
doc_id: 141950 weight: 0.0158186732814
doc_id: 161235 weight: 0.00472703733799
doc_id: 157966 weight: 0.00347568646989
doc_id: 157958 weight: 0.0030045425484

doc_id: 170332 weight: 0.00278128162755
doc_id: 159229 weight: 0.00255732999857
doc_id: 152357 weight: 0.00235706977767
doc_id: 160660 weight: 0.00210827595733
doc_id: 154115 weight: 0.00210356998421

station:
doc_id: 173088 weight: 0.00780757942152
doc_id: 142175 weight: 0.00722216464811
doc_id: 172174 weight: 0.00447184157315
doc_id: 164511 weight: 0.00285897018859
doc_id: 162036 weight: 0.00238042800545
doc_id: 163251 weight: 0.00200883350693
doc_id: 145771 weight: 0.00195173055447
doc_id: 169004 weight: 0.00171214864358
doc_id: 172759 weight: 0.00168370488385
doc_id: 165098 weight: 0.00160985908949

student:
doc_id: 155526 weight: 0.00965361471224
doc_id: 170686 weight: 0.00689259643744
doc_id: 171807 weight: 0.00653834008423
doc_id: 173651 weight: 0.00621112813818
doc_id: 170570 weight: 0.00554080378214
doc_id: 161724 weight: 0.00548665522685
doc_id: 151712 weight: 0.00547858572301
doc_id: 169136 weight: 0.00506924905285
doc_id: 161015 weight: 0.00485110921166
doc_id: 140699 weight: 0.00459541511728

location:
doc_id: 146339 weight: 0.00319753019192
doc_id: 162759 weight: 0.00099294950979
doc_id: 165098 weight: 0.000865944339933
doc_id: 170608 weight: 0.000744712132342
doc_id: 160338 weight: 0.000722698890282
doc_id: 159924 weight: 0.000703474001235
doc_id: 146169 weight: 0.000643395784645
doc_id: 158835 weight: 0.000619880936795
doc_id: 165820 weight: 0.000600574300276
doc_id: 150536 weight: 0.00058290558267

section:
doc_id: 166496 weight: 0.00242660720108
doc_id: 143425 weight: 0.00179940926141
doc_id: 170717 weight: 0.00178567572843
doc_id: 174647 weight: 0.00131244983883
doc_id: 161305 weight: 0.00128272694064
doc_id: 170652 weight: 0.00119313621711
doc_id: 149131 weight: 0.00118761011253
doc_id: 143822 weight: 0.00116374539221

doc_id: 172174 weight: 0.00114622965783
doc_id: 135916 weight: 0.00109544733121

class:

doc_id: 173512 weight: 0.00425301030433
doc_id: 159811 weight: 0.00257257855603
doc_id: 156879 weight: 0.00213260528169
doc_id: 145132 weight: 0.00210230977117
doc_id: 140976 weight: 0.00198043285321
doc_id: 162330 weight: 0.00184782502006
doc_id: 172640 weight: 0.00151231427547
doc_id: 142280 weight: 0.00149914539463
doc_id: 172078 weight: 0.00114878728126
doc_id: 167316 weight: 0.00114856246966

study:

doc_id: 160538 weight: 0.00209364038919
doc_id: 166380 weight: 0.0019779468355
doc_id: 144147 weight: 0.00169655708981
doc_id: 153807 weight: 0.00120411256003
doc_id: 173908 weight: 0.00120205072345
doc_id: 151712 weight: 0.00105625321449
doc_id: 155624 weight: 0.00102974143957
doc_id: 140282 weight: 0.00102278494013
doc_id: 169409 weight: 0.0009995220064
doc_id: 153465 weight: 0.000888792094317

science:

doc_id: 151686 weight: 0.00503769757168
doc_id: 173612 weight: 0.00451098424423
doc_id: 151712 weight: 0.00369026745219
doc_id: 166380 weight: 0.00249398599141
doc_id: 162385 weight: 0.00223640676446
doc_id: 167354 weight: 0.00199054753776
doc_id: 149128 weight: 0.00194290421312
doc_id: 154826 weight: 0.0019405129464
doc_id: 164375 weight: 0.00173038240602
doc_id: 170686 weight: 0.00159178876219

review:

doc_id: 161277 weight: 0.00143105691078
doc_id: 144482 weight: 0.00114821292556
doc_id: 164120 weight: 0.00112826691276
doc_id: 162385 weight: 0.00108505669057
doc_id: 167579 weight: 0.000998359150475
doc_id: 148051 weight: 0.000852667956683
doc_id: 157596 weight: 0.000836331960845
doc_id: 170146 weight: 0.000823808425842
doc_id: 173238 weight: 0.000761411712196
doc_id: 159599 weight: 0.000700150989536

figure:

doc_id: 174074 weight: 0.00981685747021
doc_id: 146773 weight: 0.000813323724782
doc_id: 171939 weight: 0.000771835825298
doc_id: 142323 weight: 0.000676025016786
doc_id: 155624 weight: 0.000667706234492
doc_id: 164055 weight: 0.000572086524469
doc_id: 160591 weight: 0.000517457572531
doc_id: 148869 weight: 0.000512697778828
doc_id: 149068 weight: 0.000498610875001
doc_id: 145560 weight: 0.000484373692971

author:

doc_id: 161842 weight: 0.00298429744925
doc_id: 154374 weight: 0.00149489438497
doc_id: 165627 weight: 0.00144088304938
doc_id: 146248 weight: 0.00131986938402
doc_id: 172881 weight: 0.000921024222496
doc_id: 166302 weight: 0.000906551811709
doc_id: 157426 weight: 0.000851430892815
doc_id: 166531 weight: 0.000807496836295
doc_id: 140356 weight: 0.000777533117212
doc_id: 145431 weight: 0.000707576497463

stage:

doc_id: 145166 weight: 0.00377703189966
doc_id: 151932 weight: 0.00259112064789
doc_id: 141029 weight: 0.0018327132624
doc_id: 163368 weight: 0.00155459724204
doc_id: 166781 weight: 0.00114577001202
doc_id: 165887 weight: 0.00111645526484
doc_id: 149804 weight: 0.000898304209345
doc_id: 163036 weight: 0.000855266189121
doc_id: 163618 weight: 0.000833753065303
doc_id: 167626 weight: 0.00079630030787

course degree:

doc_id: 167316 weight: 0.0038669153473
doc_id: 151712 weight: 0.00266821858369
doc_id: 143335 weight: 0.0017343420794
doc_id: 173651 weight: 0.00159719284047
doc_id: 163157 weight: 0.00150246641171
doc_id: 165478 weight: 0.001346452429
doc_id: 173741 weight: 0.00118543818705
doc_id: 170686 weight: 0.00107420154524
doc_id: 140699 weight: 0.000957405486406
doc_id: 147456 weight: 0.000957405486406

publication:

doc_id: 142539 weight: 0.00104255340728
doc_id: 148731 weight: 0.000833240247255

doc_id: 165243 weight: 0.000768197247469
doc_id: 151686 weight: 0.00076458007385
doc_id: 172652 weight: 0.000752510670801
doc_id: 154771 weight: 0.000687767859945
doc_id: 158772 weight: 0.000676133530499
doc_id: 161842 weight: 0.000654418914595
doc_id: 154374 weight: 0.000628305463347
doc_id: 149714 weight: 0.000606311106404

interview:
doc_id: 174609 weight: 0.00144175820018
doc_id: 167290 weight: 0.00129915761914
doc_id: 148170 weight: 0.00125792600205
doc_id: 145663 weight: 0.00124945128691
doc_id: 157426 weight: 0.00124475763068
doc_id: 170947 weight: 0.00124209626739
doc_id: 171305 weight: 0.000964454685033
doc_id: 151360 weight: 0.00084404798245
doc_id: 159119 weight: 0.000828064178261
doc_id: 165908 weight: 0.000811701233422

network:
doc_id: 147130 weight: 0.00622402403915
doc_id: 171742 weight: 0.00585203426422
doc_id: 173512 weight: 0.005289190389
doc_id: 159846 weight: 0.00472632636656
doc_id: 151799 weight: 0.00446520181328
doc_id: 165668 weight: 0.0043012451842
doc_id: 172174 weight: 0.00386706370163
doc_id: 154549 weight: 0.00267234526057
doc_id: 169771 weight: 0.00266926463633
doc_id: 146339 weight: 0.00254731361927

space:
doc_id: 140282 weight: 0.00550325041816
doc_id: 152664 weight: 0.00238008675751
doc_id: 165668 weight: 0.00202111099873
doc_id: 158382 weight: 0.00198256119766
doc_id: 160020 weight: 0.00183505664449
doc_id: 173181 weight: 0.00150168973032
doc_id: 167774 weight: 0.00126752674099
doc_id: 157093 weight: 0.0012427742858
doc_id: 151932 weight: 0.00121745752003
doc_id: 169321 weight: 0.00103228546687

people:
doc_id: 159229 weight: 0.00159810015106
doc_id: 162699 weight: 0.00095823848254
doc_id: 163579 weight: 0.000918149010656
doc_id: 169553 weight: 0.000637313153476
doc_id: 171192 weight: 0.000589901177638

doc_id: 154944 weight: 0.000581747307027
doc_id: 169305 weight: 0.000555113092971
doc_id: 138301 weight: 0.000553304819087
doc_id: 162279 weight: 0.000524828146052
doc_id: 159405 weight: 0.000494583812212

element:
doc_id: 172640 weight: 0.00665106047688
doc_id: 156411 weight: 0.00207035041907
doc_id: 143809 weight: 0.00137405038598
doc_id: 161487 weight: 0.00104637468862
doc_id: 152176 weight: 0.001040050446
doc_id: 168377 weight: 0.000975717428722
doc_id: 157700 weight: 0.000949157785247
doc_id: 174662 weight: 0.000838057177562
doc_id: 170717 weight: 0.00079891325762
doc_id: 169652 weight: 0.000763262827307

battle:
doc_id: 146248 weight: 0.00468065053351
doc_id: 165630 weight: 0.00443489383401
doc_id: 147394 weight: 0.00423242578757
doc_id: 160665 weight: 0.00376650286345
doc_id: 140986 weight: 0.00365604669886
doc_id: 148732 weight: 0.00339658649164
doc_id: 160664 weight: 0.00293804050614
doc_id: 140784 weight: 0.00261557943165
doc_id: 146646 weight: 0.00237126706716
doc_id: 171473 weight: 0.00236287617166

actor:
doc_id: 169212 weight: 0.00824498338832
doc_id: 169749 weight: 0.00268008073003
doc_id: 169741 weight: 0.00224975790859
doc_id: 164390 weight: 0.00199784933991
doc_id: 161436 weight: 0.00199147401363
doc_id: 166959 weight: 0.0017981179532
doc_id: 163368 weight: 0.00159821601288
doc_id: 156198 weight: 0.00152319909259
doc_id: 164237 weight: 0.00149240735555
doc_id: 169771 weight: 0.00147597053756

library:
doc_id: 150983 weight: 0.00112828736673
doc_id: 157596 weight: 0.000949011385718
doc_id: 147845 weight: 0.000927381496585
doc_id: 171807 weight: 0.000909248338831
doc_id: 160892 weight: 0.000880354037326
doc_id: 170570 weight: 0.000807218612031
doc_id: 165627 weight: 0.000758766679024
doc_id: 169136 weight: 0.000704950525829

doc_id: 144868 weight: 0.000698021955614
doc_id: 174145 weight: 0.000687457033372

value:
doc_id: 165259 weight: 0.0110089141641
doc_id: 143012 weight: 0.00747015691295
doc_id: 164055 weight: 0.00694273618724
doc_id: 172640 weight: 0.00431334720047
doc_id: 145865 weight: 0.00347702254243
doc_id: 149984 weight: 0.00271188206168
doc_id: 161305 weight: 0.00210459770915
doc_id: 168568 weight: 0.00200701342815
doc_id: 152205 weight: 0.00199186937657
doc_id: 146806 weight: 0.00162268851492

musician:
doc_id: 164003 weight: 0.00190424329854
doc_id: 165987 weight: 0.00141496751534
doc_id: 153079 weight: 0.00139513152213
doc_id: 165604 weight: 0.001195727242
doc_id: 161123 weight: 0.00117186069691
doc_id: 148397 weight: 0.00109506808421
doc_id: 160810 weight: 0.00105451000701
doc_id: 171038 weight: 0.000870114814531
doc_id: 161105 weight: 0.00085877399223
doc_id: 163725 weight: 0.000849987603519

museum:
doc_id: 166380 weight: 0.00421145673672
doc_id: 136196 weight: 0.00307283786268
doc_id: 171627 weight: 0.00286024847759
doc_id: 155125 weight: 0.00222537875666
doc_id: 172013 weight: 0.00186173766258
doc_id: 146169 weight: 0.00169211062133
doc_id: 163543 weight: 0.001541622542
doc_id: 145771 weight: 0.00152138874614
doc_id: 160774 weight: 0.0014566487995
doc_id: 137979 weight: 0.00139417067229

paper:
doc_id: 166105 weight: 0.00971604229706
doc_id: 151959 weight: 0.00222787336294
doc_id: 172985 weight: 0.00220440019837
doc_id: 162300 weight: 0.00109556033955
doc_id: 164055 weight: 0.00106664525728
doc_id: 136566 weight: 0.000958687487432
doc_id: 164602 weight: 0.00092977002232
doc_id: 161724 weight: 0.000905241495101
doc_id: 142950 weight: 0.000749795985842
doc_id: 164500 weight: 0.000722180150366

basis:

doc_id: 157093 weight: 0.00390298985846
doc_id: 172189 weight: 0.000697960747574
doc_id: 165230 weight: 0.000653087873078
doc_id: 145865 weight: 0.000591404735823
doc_id: 146248 weight: 0.000583828020346
doc_id: 156668 weight: 0.000436318281163
doc_id: 166380 weight: 0.000378444460709
doc_id: 153851 weight: 0.000377660390693
doc_id: 152664 weight: 0.000363356973903
doc_id: 151959 weight: 0.00036072047871

parent:

doc_id: 155624 weight: 0.00253164936971
doc_id: 159405 weight: 0.00133724484975
doc_id: 149154 weight: 0.00119759329288
doc_id: 171939 weight: 0.000975487711859
doc_id: 144147 weight: 0.000926896688618
doc_id: 162699 weight: 0.000704039175249
doc_id: 147965 weight: 0.000673646227247
doc_id: 150136 weight: 0.000662780965517
doc_id: 157960 weight: 0.000644165954728
doc_id: 167316 weight: 0.00064332555557

attention:

doc_id: 159599 weight: 0.000788858934783
doc_id: 170338 weight: 0.000559561596478
doc_id: 156507 weight: 0.000554929464056
doc_id: 168577 weight: 0.000533438826987
doc_id: 170344 weight: 0.000520461795481
doc_id: 154584 weight: 0.000515129195118
doc_id: 151140 weight: 0.000505123336673
doc_id: 149068 weight: 0.000481884435561
doc_id: 144930 weight: 0.000478217591409
doc_id: 141931 weight: 0.000477460678476

episode:

doc_id: 164339 weight: 0.00389781421607
doc_id: 145143 weight: 0.00385466333968
doc_id: 171063 weight: 0.00303864688999
doc_id: 168117 weight: 0.00299390368995
doc_id: 143797 weight: 0.00274743486154
doc_id: 160953 weight: 0.00272402945315
doc_id: 163252 weight: 0.00266250370701
doc_id: 161897 weight: 0.002217841663
doc_id: 154813 weight: 0.00194658573247
doc_id: 157788 weight: 0.00194218287419

foundation:

doc_id: 149128 weight: 0.00247820625723
doc_id: 146188 weight: 0.00153698873978

doc_id: 151712 weight: 0.00102965554741
doc_id: 154584 weight: 0.00085854865853
doc_id: 154374 weight: 0.000759924568372
doc_id: 171050 weight: 0.000670712505917
doc_id: 168465 weight: 0.000602115083156
doc_id: 168632 weight: 0.000568097281848
doc_id: 144503 weight: 0.000558628993817
doc_id: 152625 weight: 0.00055615718411

black board:
doc_id: 159924 weight: 0.00514449762338
doc_id: 165140 weight: 0.00178167499478
doc_id: 172985 weight: 0.00161922303611
doc_id: 148281 weight: 0.00150709867725
doc_id: 171989 weight: 0.00126672263855
doc_id: 157960 weight: 0.00108831404522
doc_id: 160664 weight: 0.000787030016565
doc_id: 167538 weight: 0.000687407410541
doc_id: 152654 weight: 0.000673665590979
doc_id: 140784 weight: 0.000656847580163

sport:
doc_id: 162411 weight: 0.00426887692686
doc_id: 152640 weight: 0.00424011644709
doc_id: 173198 weight: 0.0027697180771
doc_id: 146996 weight: 0.00225975587322
doc_id: 159932 weight: 0.0020560738901
doc_id: 159298 weight: 0.00185042044209
doc_id: 159975 weight: 0.00170808040808
doc_id: 153992 weight: 0.00151495219688
doc_id: 149282 weight: 0.00133754193788
doc_id: 167104 weight: 0.00111249519912

master:
doc_id: 148281 weight: 0.00235894821648
doc_id: 161015 weight: 0.00143589610618
doc_id: 147176 weight: 0.000946187324533
doc_id: 167354 weight: 0.00088014932409
doc_id: 165820 weight: 0.000720077672823
doc_id: 166683 weight: 0.000683820240705
doc_id: 146339 weight: 0.000638963432812
doc_id: 162300 weight: 0.000623951641318
doc_id: 166781 weight: 0.000578636717647
doc_id: 150186 weight: 0.000565782806585

health:
doc_id: 146334 weight: 0.00359069528377
doc_id: 148720 weight: 0.0028602142457
doc_id: 146717 weight: 0.00233927974319
doc_id: 168465 weight: 0.00166122129966
doc_id: 168932 weight: 0.00134933831247

doc_id: 137986 weight: 0.00128239117277
doc_id: 170584 weight: 0.00125085479138
doc_id: 156754 weight: 0.00123629214369
doc_id: 171988 weight: 0.00116452848363
doc_id: 140968 weight: 0.00114703171035

management:
doc_id: 142983 weight: 0.00145623778258
doc_id: 167354 weight: 0.00136714097739
doc_id: 150136 weight: 0.00118258153933
doc_id: 163118 weight: 0.00116651065994
doc_id: 154115 weight: 0.00114265021463
doc_id: 163327 weight: 0.000992954761886
doc_id: 158181 weight: 0.000955829076873
doc_id: 146072 weight: 0.000768285596631
doc_id: 158158 weight: 0.000759465399449
doc_id: 143215 weight: 0.000738493088183

theory:
doc_id: 152664 weight: 0.00744866407405
doc_id: 169305 weight: 0.00355429278105
doc_id: 167394 weight: 0.00308963879599
doc_id: 156411 weight: 0.00300274270485
doc_id: 173505 weight: 0.00272879706227
doc_id: 156115 weight: 0.00221897059213
doc_id: 168918 weight: 0.00214443328324
doc_id: 162385 weight: 0.0020822802051
doc_id: 152205 weight: 0.00204235648133
doc_id: 143608 weight: 0.00198151791131

season:
doc_id: 163252 weight: 0.00935770230446
doc_id: 165278 weight: 0.00560147875453
doc_id: 163215 weight: 0.00505459056051
doc_id: 170608 weight: 0.00475530134612
doc_id: 141390 weight: 0.00425016660276
doc_id: 167090 weight: 0.00405118040221
doc_id: 147041 weight: 0.00399088610427
doc_id: 161459 weight: 0.00388893340437
doc_id: 151744 weight: 0.00329947871855
doc_id: 156093 weight: 0.00310361003085

theatre:
doc_id: 163618 weight: 0.00786829947848
doc_id: 149804 weight: 0.00767508651742
doc_id: 164237 weight: 0.0050186357963
doc_id: 163563 weight: 0.00320147380064
doc_id: 163734 weight: 0.00267423097131
doc_id: 167626 weight: 0.0025447111413
doc_id: 163733 weight: 0.00251615682951
doc_id: 161051 weight: 0.00244477230984

doc_id: 161048 weight: 0.00243336147059
doc_id: 163036 weight: 0.0022206815321

journal:
doc_id: 162385 weight: 0.00243593136152
doc_id: 173908 weight: 0.00180962234971
doc_id: 167347 weight: 0.0015810131433
doc_id: 163062 weight: 0.00119395354513
doc_id: 162903 weight: 0.00106965531602
doc_id: 169409 weight: 0.00102404985681
doc_id: 169305 weight: 0.00100521824879
doc_id: 152678 weight: 0.000971851834684
doc_id: 164695 weight: 0.000970152495928
doc_id: 171807 weight: 0.00088560289293

earth:
doc_id: 164406 weight: 0.00447510763708
doc_id: 140282 weight: 0.00319051265033
doc_id: 149068 weight: 0.0028348580598
doc_id: 174609 weight: 0.00265428737053
doc_id: 160664 weight: 0.00247759839739
doc_id: 174443 weight: 0.00175698738117
doc_id: 151932 weight: 0.00167543537145
doc_id: 143608 weight: 0.00164682734357
doc_id: 143797 weight: 0.00146614436999
doc_id: 159668 weight: 0.0013718468966

morning:
doc_id: 142421 weight: 0.000978118354106
doc_id: 141958 weight: 0.000902270727293
doc_id: 172985 weight: 0.000846220576052
doc_id: 169771 weight: 0.000764552471176
doc_id: 158287 weight: 0.000750679543272
doc_id: 148702 weight: 0.000664492064479
doc_id: 154989 weight: 0.000642699171685
doc_id: 167485 weight: 0.000631793642754
doc_id: 169749 weight: 0.00062472659977
doc_id: 171807 weight: 0.000624028581223

ΜΕΡΟΣ Β'

Στο δεύτερο μέρος της εργαστηριακής άσκησης, εφαρμόσαμε την τεχνική containment για να βρούμε το ποσοστό ομοιότητας των 5 μεγαλύτερων σε tokens κειμένων της συλλογής. Απαραίτητη προϋπόθεση ήταν να αναπαραστήσουμε τα tokenized κείμενα του Α' μέρους σε shingles μορφή, μεγέθους 2. Η επεξεργασία της μετατροπής σε 2-shingles και ο υπολογισμός του containment γίνονται στο script containments.py, εκτελώντας:

containments.py

Για την 2-shingles αναπαράσταση, διαβάζονται οι γραμμές κάθε tokenized αρχείου και αφού αφαιρεθεί ο χαρακτήρας '\n' και μετατραπούν σε πεζά, γίνεται η αποθήκευση της 2-

shingles μορφής του κάθε αρχείου σε νέο με το πρόθεμα shinglezed_ . Στα 2-shingles αρχεία κάθε token αποθηκεύεται με το επόμενο του σε μία γραμμή με ένα tab να τα διαχωρίζει. Ένα δείγμα της 2-shingles αναπαράστασης του κειμένου με id 135916 είναι το παρακάτω:

bacliff ,
, texas
texas bacliff
bacliff is
is a
a census
census -

Οι 2-shingles αναπαραστάσεις των αρχείων αποθηκεύονται στον φάκελο shingles.

Για να βρούμε τα 5 μεγαλύτερα κείμενα, βρίσκουμε τον αριθμό των tokens που έχει το κάθε tokenized κείμενο και τον βάζουμε σε μία λίστα που κρατάει όνομα αρχείου και αριθμό tokens. Η λίστα ταξινομείται με βάση τον αριθμό των tokens και κρατάμε τα 5 πρώτα κείμενα.

Για κάθε ένα από τα 5 κείμενα υπολογίζουμε το σύνολο Α, δηλαδή μία λίστα με τα singles του. Επίσης διατηρούμε μία άλλη λίστα με τα signles των υπόλοιπων 4 κειμένων. Η μετρική containement ενός κειμένου είναι το πλήθος των shingles του κειμένου που επαναλαμβάνονται και στα υπόλοιπα 4 δια του αριθμού των shingles του.

Στην υλοποίηση, για κάθε κείμενο μέσα σε ένα διπλό loop, διατηρούμε και κάνουμε populate τις παρακάτω 2 λίστες:

```
# 2-shingles list of current text
A = []
# 2-shignles list for the rest of 4 texts
A_coll = []
```

Τα κοινά στοιχεία ανάμεσα στις 2 λίστες τα βρίσκουμε ως εξής:

```
# arithmos koinwn shingles sto A, A_coll
commons = list(set(A) & set(A_coll))
```

Το ποσοστό των containments των 5 μεγαλύτερων κειμένων αποθηκεύεται στο αρχείο containments.txt. Τα αποτελέσματα της επεξεργασίας είναι τα παρακάτω:

id: 169769 containment: 6.4821558630 7%
id: 152847 containment: 3.0872233587 5%
id: 165668 containment: 3.8790644609 2%
id: 163883 containment: 6.2256809338 5%
id: 160784 containment: 7.0362037226 4%

Για την εργασία:

- Χρησιμοποιήθηκε η έκδοση 2.7 της Python
- Έγινε σε περιβάλλον Windows 7 Professional