Προγραμματιστικά Εργαλεία και Τεχνολογίες για Επιστήμη Δεδομένων

Εξέταση εργαστηρίου Python, 2/12/2021, Νίκος Παπασπύρου.

Πρόβλημα "word-groups"

Σε αυτή την άσκηση υποθέτουμε ότι οι λέξεις αποτελούνται από μικρά γράμματα του λατινικού αλφαβήτου. Θα λέμε ότι δύο λέξεις ανήκουν στην ίδια "ομάδα" αν γράφονται με το ίδιο σύνολο γραμμάτων. Για παράδειγμα, οι τέσσερις λέξεις "stream", "matters", "smarter" και "smartest" ανήκουν στην ίδια ομάδα γιατί για να γραφούν χρειάζεται ακριβώς το ίδιο σύνολο 6 γραμμάτων: {a, e, m, r, s, t}.

Δίνεται ένα αρχείο κειμένου αποτελούμενου από τέτοιες λέξεις, μία σε κάθε γραμμή. Για παράδειγμα, έστω το αρχείο small.txt που περιέχει τα εξής:

```
$ cat small.txt
be
smarter
now
and
forget
the
rest
here
is
what
matters
the
street
that
crosses
the
stream
is
the
longest
of
all
streets
```

Θα ονομάζουμε φιλικότητα (friendliness) μιας λέξης το πλήθος των διαφορετικών μεταξύ τους λέξεων που ανήκουν στην ίδια ομάδα με αυτήν. Για παράδειγμα, η φιλικότητα της λέξης "street" στο παραπάνω αρχείο λέξεων είναι 3, γιατί η ομάδα στην οποία ανήκει περιέχει τρεις λέξεις: "rest", "street" και "streets". Προφανώς η φιλικότητα κάθε λέξης του κειμένου είναι τουλάχιστον 1, αφού η ίδια η λέξη ανήκει στην ομάδα της.

Ζητείται να γράψετε ένα πρόγραμμα σε Python που να δέχεται στο command line το όνομα του αρχείου λέξεων. Το πρόγραμμά σας πρέπει να υπολογίζει και να εκτυπώνει τα εξής:

1. Πόσες διαφορετικές ομάδες λέξεων υπάρχουν.
2. Πόσες λέξεις περιέχει η μεγαλύτερη ομάδα.
3. Ποιες είναι οι λέξεις με τη μεγαλύτερη φιλικότητα, σε αλφαβητική σειρά.

Ακολουθούν δύο παραδείγματα χρήσης του προγράμματος. Το δεύτερο αρχείο λέξεων είναι το medium.txt. Τα ζητούμενα (1)-(3) πρέπει να εκτυπώνονται στην οθόνη **ακριβώς** στην παρακάτω μορφή.

```
$ ./word-groups.py small.txt
15 different group(s)
largest group has 3 word(s)
friendliest word(s) are:
- matters
- rest
- smarter
- stream
- street
- streets
```

```
$ ./word-groups.py medium.txt
1063 different group(s)
largest group has 3 word(s)
friendliest word(s) are:
- appears
- else
- less
- papers
- seat
- sell
- spare
- state
- taste
```

Το πρόγραμμά σας θα εκτελεστεί με πολλά διαφορετικά αρχεία λέξεων, το μεγαλύτερο από τα οποία θα περιέχει μερικά εκατομμύρια λέξεων. Μπορείτε να θεωρήσετε ότι το μήκος των λέξεων δε θα υπερβαίνει το 50.

Προσοχή: Δείτε στην ιστοσελίδα του μαθήματος τι πρέπει να παραδώσετε και πώς!

Καλή επιτυχία!