Document Retrieval
Problem ID: p09documentretrieval

Aðvörun: Þetta er krefjandi verkefni, líklega erfiðasta forritunarverkefnið sem þið hafið unnið að hingað til í námskeiðinu. Þess vegna er enn mikilvægara nú en áður að byrja að vinna að verkefninu eins snemma og mögulegt er.

Bakgrunnur: Í skjalaheimt (e. document retrieval) eru skjöl fundin sem passa við ákveðin leitarskilyrði sem notandi setur fram. Þekktasta dæmið er vefleit þar sem notandi slær inn leitarstreng (mengi af leitarorðum) og leitarvélin finnur vefsíður sem passa við leitarstrenginn. Raunveruleg skjalaheimt getur verð nokkuð erfið þar sem taka þarf tillit til margra mismundani þátta. Í þessu verkefni munið þið hins vegar útfæra mjög einfalda leitarvél.

Lýsing: Í þessu verkefni eru skjalasöfnin geymd í textaskrám. Í lok hvers skjals í safni er ein lína sem inniheldur aðeins strenginn `"<END OF DOCUMENT>"`. Notandi forritsins slær inn nafn á textaskrá sem inniheldur skjalasafn, les inn einstök skjöl úr skránni og geymir innihald þeirra sem lista af strengjum – einn (langan) streng fyrir sérhvert skjal. Ef inntaksskráin finnst ekki þá á forritið að hætta keyrslu án þess að skrifa nokkuð út.

Til að fletta upp leitarorðum þarf forritið að halda utan um hvaða orð koma fyrir í sérhverju skjali.. Þið eigið að nota **uppflettitöflu** (e. dictionary) í þeim tilgangi. Sérhver færsla í uppflettitöflunnni skal hafa orð í lágstöfum sem **lykil** (e. key) og **gildið** (e. value) á að vera **mengi** (e. set) af númerum þeirra skjala sem orðið kemur fyrir í. Greinarmerki í upphafi og enda orðs skal fjarlægja.

Forritið leyfir notanda að framkvæma þrjár aðgerðir:

1. **Search**: Ef þessi aðgerð er valin þá er notandinn jafnframt beðinn um að slá inn leitarstreng. Forritið skal síðan prenta út númer þeirra skjala (í hækkandi röð) í skjalasafninu sem innihalda öll orðin í leitarstrengnum. Ekki er gerður greinarmunur á hástöfum og lágstöfum og röð orðanna í leitarstrengnum skiptir ekki máli. Ef ekkert skjal í safninu inniheldur öll orðin í leitarstrengnum þá skal forritið skrifa út skilaboðin `"No match"`.
2. **Print**: Ef þessi aðgerð er valin þá er notandinn jafnframt beðinn um að slá inn númer skjals. Forritið skal síðan prenta út innihald viðkomandi skjals. Ef ekkert skjal í safninu passar við númerið sem slegið var inn þá skrifar forritið út skilaboðin `"No match"`.
3. **Quit**: Ef notandinn velur hvorki aðgerð 1 né 2 þá hættir forritið keyrslu.

Forritið skal bjóða notandanum endurtekið upp á að framkvæma aðgerðirnar að ofan þangað til að hann velur að hætta.

Dæmi: Dæmi um skjalasafn er skráin `example.txt`:

```
This is an example document collection file.
It contains three documents.
This is text from the first document.
<END OF DOCUMENT>
This is text from the second document. Each document is
of variable length.
<END OF DOCUMENT>
And now we are in the third document, which is indeed
the "largest" one
measured in the length of the
text that appears in the document.
<END OF DOCUMENT>
```

Inntak

Fyrsta línan í inntakinu inniheldur nafn á skrá sem geymir texta úr einu eða fleiri skjölum. Síðan endurtekið, í línu $2i$, þar sem $1 \leq i \leq 5$, er inntakið eitt af eftirtöldu:

- Strengurinn `"search"`: Í þessu tilfelli geymir lína $2i + 1$ runu af einum til þremur strengjum (leitarorðum) aðskildum með hvítum bilum.
- Strengurinn `"print"`: Í þessu tilfelli geymir lína $2i + 1$ eina heiltölu sem stendur fyrir skjalið sem á að prenta út.
- Strengurinn `"quit"`: Í þessu tilfelli fylgir ekkert meira inntak.

Úttak

Þegar ekki er hægt að opna inntaksskrána þá skrifar forritið ekki neitt út. Annars, í sérhverri ítrun, er inntakið:

- Svar fyrir leitaraðgerð er annað hvort:
  - Strengurinn `"Documents matching search: "` og þar á eftir runa af númerum (í hækkandi röð) þeirra skjala sem innihalda öll orð leitarstrengsins,
  - eða strengurinnn `"No match"` ef ekkert skjal passar við leitarstrenginn.
- Svar við útprentunaraðgerð er annað hvort:
  - Strengurinn `"Document number {i}:"`, þar sem i er númer þess skjals sem beðið var um að prenta út og þar á eftir (í næstu línu) fylgir texti skjals númer i,
  - eða strengurinn `"No match"` ef ekkert skjal í safninu passar við hið innslegna númer i.

Stigagjöf

Sérhvert falið prufutilvik veitir 5 stig og eru þau 20 talsins. Ábendingar fylgja prufutilvikunum, sem þú sérð ef þú færð rangt á þeim, og lýsa þær hvað er verið að prófa.

Sample Input 1

```
example.txt
print
2
search
largest
print
3
print
1
search
document
quit
```

Sample Output 1

```
Document #2:
This is text from the second document. Each document is
of variable length.
Documents matching search: 3
Document #3:
And now we are in the third document, which is indeed
the "largest" one
measured in the length of the
text that appears in the document.
Document #1:
This is an example document collection file.
It contains three documents.
This is text from the first document.
Documents matching search: 1 2 3
```

Sample Input 2

```
example.txt
print
4
search
document from
search
the text document
search
programming
print
3
quit
```

Sample Output 2

```
No match
Documents matching search: 1 2
Documents matching search: 1 2 3
No match
Document #3:
And now we are in the third document, which is indeed
the "largest" one
measured in the length of the
text that appears in the document.
```