Puzzle pieces

We have cut up a picture into narrow vertical strips. These image strips are located somewhere on the Internet, each with their own URL (uniform resource locator). Your mission (*should you decide to accept it*) is to find the URLs and download all image stripes to re-create the original image.

The slice URLs are hidden inside apache log files (the open source apache web server is the most widely used server on the internet). Here is what a single line from the log file looks like (this really is what apache log files look like):

```
212.77.55.128 - - [06/Feb/2010:00:24:35 -0700] "GET spoj/problems/puzzle/eukuk-ruwpbl.jpg HTTP/1.1" 404 493104 "-" "googlebot-mscrawl-moma (enterprise; bar-XYZ; foo123@google.com,fo
128.34.153.184 - - [06/Feb/2010:00:26:12 -0700] "GET uypys/bclkvy/tgtquo-rds-zmcps HTTP/1.0" 301 858367 "-" "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.1.6; Google-TR-5.1.
163.196.162.170 - - [06/Feb/2010:00:33:25 -0700] "GET spoj/problems/puzzle/jwbsb-zcckx.jpg HTTP/1.0" 200 453671 "-" "Mozilla/5.0 (Macintosh; U; PPC Mac OS X Mach-O; en-US; rv:1.8.0.1
174.76.115.140 - - [06/Feb/2010:00:36:17 -0700] "GET spoj/problems/puzzel/omkzx-idk.jpg HTTP/1.1" 200 855063 "-" "googlebot-mscrawl-moma (enterprise; bar-XYZ; foo123@google.com,foo1
5.254.189.77 - - [06/Feb/2010:00:46:03 -0700] "GET bgcic/cdx/nqegmo/pzdvszjanh-jhqbj-fxqbfw-wzb HTTP/1.0" 200 320132 "-" "Mozilla/5.0 (Macintosh; U; PPC Mac OS X Mach-O; en-US; rv:1
```

The first few numbers are the IP-address of the requesting browser. The most interesting part is the "GET *path* HTTP" showing the path of a web request received by the server. The path itself never contains spaces, and is separated from the GET and HTTP by spaces.

In the log file, the image strips have a path name that includes the term puzzle. URLs that occur multiple times, are to be deduplicated. The file name of each image strip contains one or more hyphens (-). The image strips must be arranged alphabetically according to the part after the last hyphen. No distinction should be made between uppercase and lowercase letters. The sort key of the URL on the first line in the example apache file for example is ruwbpl.jpg.

Assignment

Write a function puzzlepieces to which the location of a log file should be passed as an argument. In that log the strips of the image are hidden. The function should return a list with the URLs of the image strips, deduplicated and sorted according to the key that was discussed above.

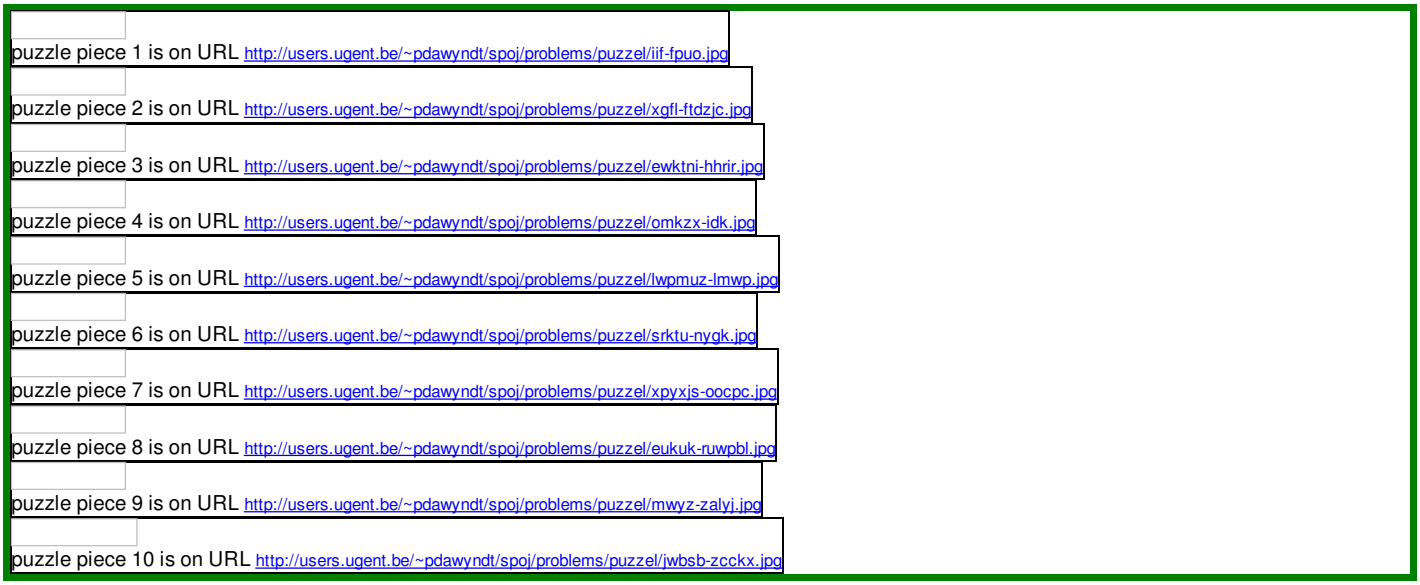
Note: By way of verification, you also get to see a graphical representation of the image strips that you have achieved from the log upon feedback.

Example

In the following interactive session we assume that the file angkorwat.log is in the current directory.

```
>>> puzzlepieces('angkorwat.log')
['spoj/problems/puzzle/iif-fpuo.jpg', 'spoj/problems/puzzle/xgfl-ftdzjc.jpg', 'spoj/problems/puzzle/ewktni-hhrir.jpg', 'spoj/problems/puzzle/omkzx-idk.jpg', 'spoj/problems/puzzle/lwpmuz-lmwp.jpg', 's
```

If we collect these images from the web server http://users.ugent.be/~pdawyndt and put them next to each other, then the strips form the image below:



We hebben een afbeelding verknipt in smalle verticale stroken. Deze afbeeldingsstroken bevinden zich ergens op het Internet, elk met hun eigen URL (uniform resource locator). Je opdracht (*should you decide to accept it*) bestaat erin de URLs van de afbeeldingsstroken te achterhalen en er de originele afbeelding mee te reconstrueren.

De URLs van de afbeeldingsstroken werden verborgen in een apache logbestand (de *open source* apache webserver is de meest gebruikte server op het Internet). Onderstaand voorbeeld geeft aan hoe de informatie in een apache logbestand er uitziet:

```
212.77.55.128 - - [06/Feb/2010:00:24:35 -0700] "GET spoj/problems/puzzel/eukuk-ruwpbl.jpg HTTP/1.1" 404 493104 "-" "googlebot-mscrawl-moma (enterprise; bar-XYZ; foo123@google.com,fo
128.34.153.184 - - [06/Feb/2010:00:26:12 -0700] "GET uypys/bclkvy/tgtquo-rds-zmcps HTTP/1.0" 301 858367 "-" "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.1.6; Google-TR-5.1.
163.196.162.170 - - [06/Feb/2010:00:33:25 -0700] "GET spoj/problems/puzzel/jwbsb-zcckx.jpg HTTP/1.0" 200 453671 "-" "Mozilla/5.0 (Macintosh; U; PPC Mac OS X Mach-O; en-US; rv:1.8.0.1
174.76.115.140 - - [06/Feb/2010:00:36:17 -0700] "GET spoj/problems/puzzel/omkzx-idk.jpg HTTP/1.1" 200 855063 "-" "googlebot-mscrawl-moma (enterprise; bar-XYZ; foo123@google.com,foo1
5.254.189.77 - - [06/Feb/2010:00:46:03 -0700] "GET bgcic/cdx/nqegmo/pzdvszjanh-jhqbj-fxqbfw-wzb HTTP/1.0" 200 320132 "-" "Mozilla/5.0 (Macintosh; U; PPC Mac OS X Mach-O; en-US; rv:1
```

De eerste reeks getallen stelt het IP-adres voor van de computer die een vraag heeft gericht aan de webserver. Het deel "GET *pad* HTTP" is voor ons het meest interessant, omdat daarin het pad voorkomt van de URL waarvoor een vraag gericht werd aan de webserver. Dit pad bevat zelf nooit spaties, en wordt gescheiden van GET en HTTP door spaties.

In het logbestand hebben de afbeeldingsstroken een padnaam die de term puzzel bevat. URLs die meerdere keren voorkomen, moeten ontdubbeld worden. De bestandsnaam van elke afbeeldingsstrook bevat één of meer koppeltekens (-). De afbeeldingsstroken moeten alfabetisch gerangschikt worden op basis van het gedeelte na het laatste koppelteken. Hierbij mag geen onderscheid gemaakt worden tussen hoofdletters en kleine letters. De sorteersleutel van de URL op de eerste regel in het voorbeeld apache bestand is dan bijvoorbeeld ruwbpl.jpg.

Opgave

Schrijf een functie puzzelstukken waaraan de locatie van een logbestand als argument moet doorgegeven worden. In dat logbestand zitten de stroken van een afbeelding verborgen. De functie moet een lijst met de URLs van de afbeeldingsstroken teruggeven, ontdubbeld en gesorteerd volgens de sleutel zoals die hierboven werd besproken.

Opmerking: Bij wijze van controle krijg je bij de feedback ook een grafische weergave te zien van de afbeeldingsstroken die je uit het logbestand hebt gehaald.

Voorbeeld

Bij onderstaande interactieve sessie gaan we ervan uit dat het bestand angkorwat.log zich in de huidige directory bevindt.

```
>>> puzzelstukken('angkorwat.log')
['spoj/problems/puzzel/iif-fpuo.jpg', 'spoj/problems/puzzel/xgfl-ftdzjc.jpg', 'spoj/problems/puzzel/ewktni-hhrir.jpg', 'spoj/problems/puzzel/omkzx-idk.jpg', 'spoj/problems/puzzel/lwpmuz-lmwp.jpg', 's
```

Als we deze afbeeldingen afhalen vanaf de webserver http://users.ugent.be/~pdawyndt en naast elkaar plaatsen, dan vormen de stroken onderstaande afbeelding: