Indexing Web Pages

Program File: `webpage.pas, webpage.c, or webpage.cpp`
Input Files: `webpage.in, index.htm, and others`
Output File: `webpage.out`

Write a program to create an index of a small collection of World Wide Web pages. Each "page" is a text file in a special format called HTML (HyperText Markup Language). The HTML format includes regular text and special HTML commands, which are always enclosed in angle braces. For example, the string `<A HREF="layout.htm">` is an HTML command meaning that the following text should be highlighted; a user click on the highlighted text would cause a web browser to fetch and display the file `layout.htm`.

Your program's job is to read an HTML file called `index.htm` and all the files referenced within `index.htm` by the HREF command and all the files referenced by those files, and so on until there are no new files to read. Your program should also read the file `webpage.in` containing a list of words and show a list of all the files referenced from `index.htm` which contain each word (see the Sample Output).

Assumptions

1. Any opening angle bracket (the "<" character) will be followed sooner or later by a matching closing angle bracket (the ">" character).
2. A word is any string of characters found in a file that:
   - does not lie between matching angle brackets
   - contains only letters (no spaces, hyphens, apostrophes, etc.)
   - is not part of a longer word (e.g., in "balloon", we would not consider "loon" to be a word).
3. Words will have at most 25 characters.
4. Words which differ only in case should be considered to be the same. Thus, "Word", "word", "WORD", and "wOrD" would be considered to be the same word.
5. The only HTML command you need to worry about is the HREF command, and you can assume that it will always be in the form `<A HREF="filename">`, with no additional spaces or other characters; that the name of the file is legal and in the same directory as the file you are already reading; and that the name of the file will not exceed twelve characters in length. Filenames will always end with "`.htm`".
6. HTML files may be mutually referential or self referential, but there will be at most one hundred different files to read.

Input

The initial HTML file you should start indexing will be named `index.htm`. The words in `webpage.in` will be placed one word per line, with no additional spaces.

Output

List each word in the standard input file, followed by a list of the file names it is found in, in the following format:

```
"word" can be found in the following pages:
     filename1
     filename2

"word" can be found in the following pages:
     filename3

"word" can not be found in any page.
```

Where `word` is the word from the input file, and `filename1`, `filename2`, and so on, are the names of the files containing the word. Each file name should be indented five spaces; a single blank line should separate each listing.

Sample Input

Three input files are given here: `index.htm`, `layout.htm`, and `webpage.in`.

`index.htm`

```
<HTML>
<HEAD>
<TITLE>Indexing Web Pages</TITLE>
</HEAD>

<BODY>
<P>Write a program to create an index of a small collection
of World Wide Web pages.  Each "page" is a text file in a
special format called HTML (HyperText Markup Language).  The
HTML format includes regular text and special HTML commands,
which are always enclosed in angle braces.  For example, the
string <A HREF="layout.htm"> is an HTML command meaning that
the following text should be highlighted; a user click on
the highlighted text would cause a web browser to fetch and
display the file layout.htm.</P>

<H1>Following Links</H1>
<P>Don't forget that links can be <A HREF="index.htm">
self-referential</A>!</P>
</BODY>
</HTML>
```

layout.htm

```
<A bunch of gibberish and a word>
Note that there is no rule that the file needs to be legal
HTML (if you know the rules), or that words really be
wordseiwlaoieu;a.  <A HREF="index.htm">Watch out for mutual
references!
</HTML>
```

webpage.in

```
file
index
html
HTML
recursion
word
is
```

Sample Output

```
"file" can be found in the following pages:
     index.htm
     layout.htm

"index" can be found in the following pages:
     index.htm

"html" can be found in the following pages:
     index.htm
     layout.htm

"HTML" can be found in the following pages:
     index.htm
     layout.htm

"recursion" can not be found in any page.

"word" can not be found in any page.

"is" can be found in the following pages:
     index.htm
     layout.htm
```