Thesaurus

Time limit: *1000 ms*
Memory limit: *256 MB*

IEEE Xplore

Warm greetings to all IEEEXtreme Participants from the Xplore API Team!

In this challenge, which is described below, you will be tasked with a programming challenge that uses documents in t form retrieved from the IEEE Xplore API.

For a full dynamic database search IEEE Xplore API is available for your IEEE research needs. Xplore API provides metadata on 4.9mm academic works and is now delivering full-text content on 50k 'Open Access' articles. Xplore API make your research needs fast and easy. The Xplore API Portal supports PHP, Python and Java as well as providing output in Json and XML formats. Many API use cases are listed within the API Portal.

Xplore API registration is free. To learn more about IEEE Xplore API please visit developer.ieee.org/ and register for an API key TODAY!

Challenge

An article can contain multiple terms from a hierarchical thesaurus. In this exercise the topic will be considered any te at the top level of the thesaurus. Identify all the thesaurus terms, and the topics to which they belong.

Standard input

The input will contain an XML thesaurus followed by text.

The thesaurus will be an XML document beginning with a root `<Thesaurus>` element on a line by itself. The `<Thesaurus>` element will consist of `<TermInfo>` elements, as in the following:

```
1  <Thesaurus>
2  <TermInfo>
3  <T>Power
       electronics</T
       >
4   <NT>Power
        integrated
        circuits</NT>
5  </TermInfo>
6  <TermInfo>
7   <T>Power
        integrated
        circuits</T>
8   <BT>Power
        electronics
```

```
            electronics
            </BT>
 9    <BT>Integrated
            circuits</BT>
10    <NT>Air traffic
            control</NT>
11    <UF>Integrated
            circuit
            supply</UF>
12  </TermInfo>
13  ...
14  </Thesaurus>
15
```

The `<TermInfo>` tag is the start of a new term node. `<BT>` are broader terms. For example, term `Power integrated circuits` has two broader terms `Power electronics` and `Integrated circuits`. `Power integrated circuits` will show up as narrower terms `<NT>` under both nodes. Use for terms `<UF>` are synonyms for the term `<T>` and will nee to be treated the same as the term `<T>`.

Whenever a term appears, whether as a `T` or `UF` element , you will need to use the broader terms to follow the term to a topic that has no broader terms. Since a term can have multiple broader terms `BT` , it is possible that the term belongs to multiple topics. You must count the times a term belonging to each top-level topic appears.

Text can have punctuation appended to terms. You will need to remove trailing commas, periods, question marks and exclamation points. (No other characters, including new line characters, should be removed from the text).

Lastly, the case is unimportant within the text of the body. `Space Station` , `Space station` , `space station` , etc. are matches for `<T>Space Station</T>`.

The remaining lines after the closing `</Thesaurus>` tag will be the article text.

Standard output

For each top-level topic, output `[topic] = [count]` , where `[topic]` is the top-level topic as it appears in the thesaurus, and `[count]` is the number of times the topic, or a narrower term, appears in the text. Order the output w the highest counts first. If two top-level topics have the same count, order them in alphabetical order.

Top-level topics that do not appear in the text should appear in the output as `[topic] = 0`.

Constraints and notes

- The size of input files will be less than 700 KB.
- `<T>` elements will not be duplicated in different TermInfo elements.

Input

```
<Thesaurus>
    <TermInfo><T>Power electronics</T><NT>Power integrated circuits</NT></TermInfo>
    <TermInfo><T>Power integrated circuits</T><BT>Power electronics</BT><NT>Air traffic control</NT
    <TermInfo><T>Air traffic control</T><BT>Power integrated circuits</BT></TermInfo>
    <TermInfo><T>Product safety engineering</T><NT>Consumer protection</NT></TermInfo>
    <TermInfo><T>Consumer protection</T><BT>Product safety engineering</BT></TermInfo>
    </Thesaurus>
    This article has a lot of air traffic control information.  It also talks about integrated circu
```