Search for most occurring patterns in a non language text file

Asked 13 years, 10 months ago Modified 7 years, 7 months ago

Viewed 2k times

4

I'm not completely sure this answer belongs here but I'm looking to find patterns into an ascii file.

The file itself is composed of alphanumeric characters and I want to just check for repeating patterns in the file, disregarding of separators and disregarding of natural language words or meaning, just get the most used repeated sequences.

I don't seem to find any program already developed that can do just that (as all seem to work with words, not just sets of characters). Do you know of any application that can do that?

If there's not such an application, how would you recommend I approach at coding one?

binary-data

Share

Improve this question

Follow

edited Apr 26, 2017 at 16:48

Cœur

38.6k ● 26 ● 202 ● 276

asked Feb 21, 2011 at 14:07

Jorge Córdoba
52.1k ● 11 ● 82 ● 130

Unless there's something else not stated here, this is a programming question quite suitable for SO. You do need to state what you mean by "repeating pattern". Would it be a sequence like "aaaa" or "abcabc", or would it be a string that recurs frequently within the text? In the latter case the problem is merely one of identifying and counting patterns; the biggest coding challenge (which is trivial, really) is to adopt an efficient data structure (such as an appropriate hash table). In the former case the challenge is to recognize "patterns". – whuber Feb 21, 2011 at 14:19

Is the latter case, that is, discover a sequence than recurs frecuently... I'd move it to SO but I'm not sure about how... – Jorge Córdoba Feb 21, 2011 at 14:23

@Jorge I voted to close, not because this is a bad question--it's perfectly fine--but because that initiates the process of migration. Good luck! – whuber Feb 21, 2011 at 14:38

2 Answers

Sorted by: Highest score (default)

1

I'm not aware of any existent program to do it, so I can only recommend coding solution. You will have to build a bit modified Trie with counter of occurrences on its leafs. Then the task becomes trivial: from all leafs find one with the max counter; path from the root to this leaf will be a subsequence (pattern) you searches for.

Also FYI: Longest common substring problem

(I know this question is for SO and my answer must be a comment, but I just haven't enough reputation to leave comments.)

Share Improve this answer

Follow

answered Feb 21, 2011 at 14:45

ffriend

28.4k ● 13 ● 92 ● 135

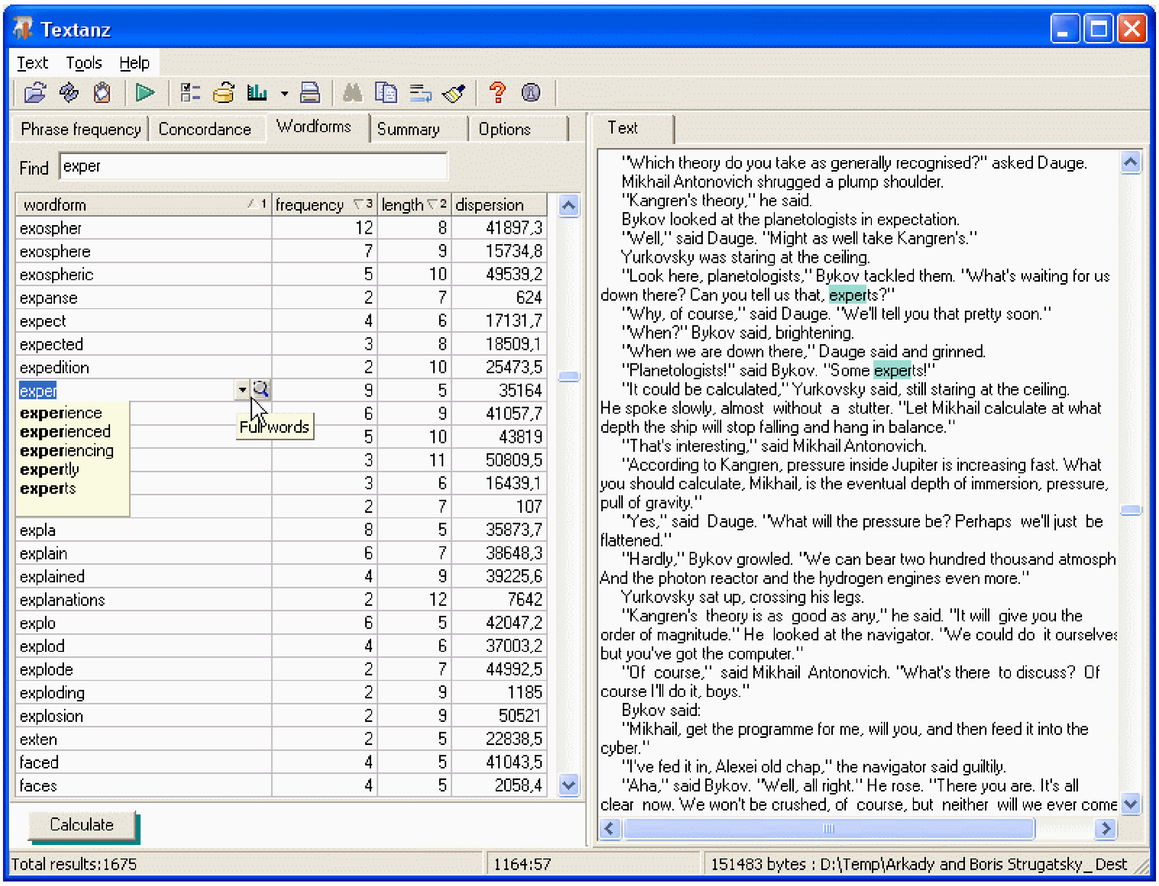
I think the question first needs clarification and refinement. After all, if a string *s* appears in the text *n* times, then each character of *s* is guaranteed to appear at least *n* times also--and is likely to occur more often than that. Thus, the problem as currently stated is solved by counting the occurrences of letters, then of digraphs, then of trigraphs, etc., and will (in any realistic case) terminate after the first step and output the single letter that occurs most frequently. – whuber Feb 21, 2011 at 16:16

@whuber: agree, but again: due to my low reputation I couldn't leave a comment, so I tried to give the most common answer. – ffriend Feb 21, 2011 at 16:51

@whuber that's exactly what I'm looking for in fact. Let's say you introduce a minimum lenght and a maximum length and it will give you the frequency for each string s.
– Jorge Córdoba Feb 22, 2011 at 8:42

1

After some searching I finally found Textanz which analyses the text and gives you a frequency count and a distribution pattern for most repeating substrings.



wordform	frequency	length	dispersion
exospher	12	8	41897,3
exosphere	7	9	15734,8
exospheric	5	10	49539,2
expanse	2	7	624
expect	4	6	17131,7
expected	3	8	18509,1
expedition	2	10	25473,5
exper	9	5	35164
	6	9	41057,7
	5	10	43819
	3	11	50809,5
	3	6	16439,1
	2	7	107
expla	8	5	35873,7
explain	6	7	38648,3
explained	4	9	39225,6
explanations	2	12	7642
explo	6	5	42047,2
explod	4	6	37003,2
explode	2	7	44992,5
exploding	2	9	1185
explosion	2	9	50521
exten	2	5	22838,5
faced	4	5	41043,5
faces	4	5	2058,4

"Which theory do you take as generally recognised?" asked Dauge.

Mikhail Antonovich shrugged a plump shoulder.

"Kangren's theory," he said.

Bykov looked at the planetologists in expectation.

"Well," said Dauge. "Might as well take Kangren's."

Yurkovsky was staring at the ceiling.

"Look here, planetologists," Bykov tackled them. "What's waiting for us down there? Can you tell us that, experts?"

"Why, of course," said Dauge. "We'll tell you that pretty soon."

"When?" Bykov said, brightening.

"When we are down there," Dauge said and grinned.

"Planetologists!" said Bykov. "Some experts!"

"It could be calculated," Yurkovsky said, still staring at the ceiling. He spoke slowly, almost without a stutter. "Let Mikhail calculate at what depth the ship will stop falling and hang in balance."

"That's interesting," said Mikhail Antonovich.

"According to Kangren, pressure inside Jupiter is increasing fast. What you should calculate, Mikhail, is the eventual depth of immersion, pressure, pull of gravity."

"Yes," said Dauge. "What will the pressure be? Perhaps we'll just be flattened."

"Hardly," Bykov growled. "We can bear two hundred thousand atmosph
And the photon reactor and the hydrogen engines even more."

Yurkovsky sat up, crossing his legs.

"Kangren's theory is as good as any," he said. "It will give you the order of magnitude." He looked at the navigator. "We could do it ourselve
but you've got the computer."

"Of course," said Mikhail Antonovich. "What's there to discuss? Of course I'll do it, boys."

Bykov said:

"Mikhail, get the programme for me, will you, and then feed it into the cyber."

"I've fed it in, Alexei old chap," the navigator said guiltily.

"Aha," said Bykov. "Well, all right." He rose. "There you are. It's all clear now. We won't be crushed, of course, but neither will we ever come

Total results:1675 | 1164:57 | 151483 bytes : D:\Temp\Arkady and Boris Strugatsky_ Dest

Share Improve this answer

Follow

answered Feb 22, 2011 at 9:01

Jorge Córdoba

52.1k ● 11 ● 82 ● 130