

Text Processing

3/30/2009

Opening Discussion

- Minute Essay comments
 - Mad libs program.
 - I know of more optimization problems. For spreadsheets I need ideas.
 - Data encryption.
 - Games conflict with current projects.
- Have you ever had things you wanted to do with text that your tools weren't good for and it took a long time?

The Problems

- I would like us to look at two different problems this week.
 - I have a bad habit of falling into a rut in my word usage. I want to count word frequency to see how often different words occur.
 - Annotating large amounts of text for things like web pages.
- The goal is to take advantage of something computers are really good at, processing large amounts of information.

Counting Words

- Today I want us to do the word counting problem.
- Obviously we need to read in the file. We can read it word by word. Each time we see a word we increment a count for it.
- To keep our counts we will use a new data structure called a map.

Maps

- Arrays and lists let us associate data with integer locations.
- Maps are more flexible. You get can associate any type of data with any type of key.
- We will use the `java.util.HashMap` for this.
- We want to have Integer counts associated with String keys.
- Let's look at the API page.

Writing the Code

- Let's write the code that will count the words.
- After we have counted them we need to print out each word and how many times it occurred.

Minute Essay

- Can you think of any other uses for a map?
Do you have any questions about what we did today?
- Remember that the project is due on Wednesday.