Programming Practice

Programming Puzzles and Competitions
CIS 4900 / 5920
Spring 2009
Next Contest

• TopCoder SRM 434
• (TOMORROW) Saturday, February 7th, 2009 @ 12:00pm
Current contest: Marathon Match 49

• Started Wednesday, February 4\textsuperscript{th}, 2009 @ 12:00pm
• Open until Wednesday, February 18\textsuperscript{th}, 2009 @ 12:00pm
Current contest: Marathon Match 49

- Problem statement:
  http://www.topcoder.com/longcontest/?module=ViewProblemStatement&compid=8907&rd=13709
Problem: Text Segmentation
Text Segmentation

- Some languages, such as Chinese and Japanese, are written without spaces between the words
Text Segmentation

• Some languages, such as Chinese and Japanese, are written without spaces between the words

• This makes it difficult for search engines to decipher the meaning of a search (and therefore to return appropriate results)
Text Segmentation

- Given a group of words with no spaces (i.e. a string of letters), what is the best segmentation of the string into words?

- Example:
  "upordown" → "up or down"
Text Segmentation

• Ambiguities are possible, e.g.: “theyouthevent” → ?
Text Segmentation

• There are at least three ways to segment “theyouthevent” into valid words:
  - “they out he vent”
  - “the you the vent”
  - “the youth event” (most likely)
Text Segmentation

• Why is this important?
Text Segmentation

• Why is this important?
• Even in English we run into juxtapositions of words (such as domain names, for example)
Text Segmentation:
Falls from grace

These examples are borrowed from Peter Norvig.
Text Segmentation: Falls from grace

- whorepresents.com

These examples are borrowed from Peter Norvig.
Text Segmentation: Falls from grace

• whorepresents.com
• therapistfinder.com

These examples are borrowed from Peter Norvig.
Text Segmentation: Falls from grace

- whorepresents.com
- therapistfinder.com
- speedofart.com

These examples are borrowed from Peter Norvig.
Text Segmentation: Falls from grace

- whorepresents.com
- therapistfinder.com
- speedofart.com
- expertsexchange.com

These examples are borrowed from Peter Norvig.
Text Segmentation:
Falls from grace

• whorepresents.com
• therapistfinder.com
• speedofart.com
• expertsexchange.com
• penisland.com

These examples are borrowed from Peter Norvig.
Text Segmentation

- How can this be done?
Text Segmentation

- \( P_0(x) = \) probability of encountering \( x \) as a single word
- \( P(y = [y_1y_2\ldots y_n]) = \) probability of encountering \( y \) as a sequence of (concatenated) words \( y_1y_2\ldots y_n \)
- \( P_{\text{max}}(y) = \) maximum probability of any segmentation of \( y \)
Text Segmentation

• We are trying to find the \([y_1 y_2 \ldots y_n]\) such that:

\[ P(y=[y_1 y_2 \ldots y_n]) = P_{\text{max}}(y) \]

• This is clearly the most likely segmentation of the word
Text Segmentation

• A simple recursion:

$$P_{\text{max}}(y) = \max_i P_0(y[0:i]) \times P_{\text{max}}(y[i:n])$$

where \( n = \text{length}(y) \)
Text Segmentation

\[ P_{\text{max}}("\text{theyouthevent}"") = \]
\[ \max( \]
\[ P_0("t") \times P_{\text{max}}("\text{heyouthevent}"") , \]
\[ P_0("\text{th}") \times P_{\text{max}}("\text{eyouthevent}"") , \]
\[ ... \]
\[ P_0("\text{theyouthevent}"") \times P_{\text{max}}(""") \]
\);
Options...

- Work on Marathon Match 49
  - Go to http://www.topcoder.com/tc

- Work on Text Segmentation
  - Go to http://ww2.cs.fsu.edu/~rosentha/icpc_practice/