An OO case study: KWIC

- Key Word in Context (KWIC) also called permuted index or concordance

- The sun ALSO rises
  the moon is a HARSH mistress
  the moon is a harsh MISTRESS
  the sun also RISES
  The SUN also rises

- List all lines or all titles or all ... ordered by keyword
  > Storage issues?
  > Specification issues?
  > Other issues?

Storage examples

- The online version of the bible we use has
  > 822,899 words, but only 17,699 are different
  > Total of 4.96 million characters in file but
    • Storing all strings requires 6.74 Mbytes, why?
    • Total storage for 822,899 pointers = 3.29 Mbytes

- Idea: store all words, but store them once and use a word
  proxy or standin as the word – the proxy doesn’t repeat the
  same word, but uses a hidden pool/dictionary of words
  > How do the real words get stored in the pool?
  > How does the proxy serve as a standin?
  > Why is this called the ‘flyweight’ pattern?

The Flyweight Pattern

- A flyweight is a shared object that can be used in multiple
  contexts simultaneously
  > Intrinsic state stored in flyweight, doesn’t depend on use
  > Extrinsic state part of context which is passed as needed
    - In some cases subclassing provides alternative to passing
  > Logically a different object exists for each occurrence, but
    physically a shared object is used for duplicates
- Think characters in a document in different fonts/styles
- Use flyweight when
  > Storage costs are high because there are lots of objects
  > Most state is extrinsic, shared objects work once extrinsic
    state is removed
  > Object identity doesn’t matter (objects in different contexts
    appear identical using ==)

Consequences of Flyweight

- Storage savings can be at the expense of runtime cost
  > Must compute and pass extrinsic context/state
  > If extrinsic context/state

- Consider the KWIC problem
  > Each ‘word’ can be a flyweight
  > What is a line?
  > What is the extrinsic state of a ‘word’ in a line?
  > How is a line printed?

- What about a concordance
  > every word and the surrounding n words are printed