HW # 2 discussion

Adam Meyers
New York University
Basics

• Most people who did this used shell scripts
• Not everyone remembered to make their script executable
• It is best to make the script of the form:
  – `grep -E 'your_regexp' $1`
• This gives me the flexibility of applying it to any file I want and also to pipe the output to any UNIX command I want (e.g., less)
Names Prefixed by “Dr.”

- Most people who did this one got good results – this was the easiest question
- It was possible for the system to overgenerate if your regexp did not represent that Dr\. is followed by a capitalized word.
- The following simple script works pretty well:
  - grep -E 'Dr\. [A-Z]' $1
Name + like + plural

- Most people who did this did OK, but it was difficult to get a really “good” result.
- The best that it was really possible to do was to get good recall (get the applicable cases) and probably overgenerate somewhat
  - One problem is that “like” is ambiguous – it is a preposition, as well as a verb.
    - Replacing 'lik(e|es|ed|ing)' with 'lik(es|ed|ing)' improves precition and has very little effect on recall
  - A second problem is that not all words ending in -s are plural
    - One student piped the result to grep -v and removed lines that contained common non-plurals ending in -s
- Sample correct answer:
  - grep -E '[A-Z][a-z]+ (lik(e|es|ed|ing) [a-z]+s([^[a-z]$)))' $1
Telephone Numbers

• Sequences of numbers in a particular pattern, separated by a hyphen or a space.
• One possible solution is as follows:
  – grep -E '([^0-9a-zA-Z])\[0-9\]{3}[- ]\[0-9\]{4}(\^[^0-9a-zA-Z])' $1
  – The preceding (^[0-9a-zA-Z]) ensures that the sequence either begins a line or is preceded by a non-number (e.g., a space)
  – The following ($[^0-9a-zA-Z]) ensures that the sequence either ends the line or is followed by a non-number.
• Most people who attempted his produced highly accurate results, but missed some examples (recall was more difficult than precision)
• Common errors:
  – requiring area codes
  – not allowing some format variation (hyphens or spaces)