LANGUAGE, FSA & REGEX

NATURAL LANGUAGE PROCESSING - CS 322.00

Blake Howald

September 15, 2017

AGENDA

- Regular Expressions
- Accuracy, Precision, Recall
- Assignment 1

REGEX

- Any FSA can be described by a Regular Expression (REGEX)
  - And any REGEX can implement a FSA
- String matching (any sequence of characters)
  - Single strings, mentions to larger patterned sequences (Temporal Expressions, Email Addresses, Phone Numbers)
- Common technique in NLP and Information Retrieval (IR)

REGEX – PRACTICAL CONSIDERATIONS

- Matches first occurrence (in Unix, Perl, Python, etc.)
- Case sensitive
  - BoJack matches 'BoJack', no 'boJack'
- Disjunction via [ ]
  - [Bb]oJack matches 'BoJack' or 'boJack'
- Range
  - [abcdefghijklmnopqrstuvwxyz] = [a-z]
  - [23456] = [2-6]

REGEX – PRACTICAL CONSIDERATIONS

- Negation [^]
  - [^b]oJack matches 'coJack', 'soJack'
  - [b^]oJack matches 'boJack', '^oJack'
- Special Characters
  - ? – preceding character or nothing
    - BoJacks? matches 'BoJacks' or 'BoJack'
  - Kleene* - zero or more occurrences
  - Kleene+ - one or more occurrences
  - . – wildcard (match any example)

REGEX – PRACTICAL CONSIDERATIONS

- Anchors
  - ^ - line start
  - $ - line end
  - \b, \B
- Precedence and Grouping
  - | & () – (B|C)oJack is the same as BoJack|CoJack but not B|CoJack
- Occurrence
  - {} - 'Hollywood' = 'Hol{2}ywo{2}d'

REGEX – ADDITIONAL CONSIDERATIONS

- Shortcuts – see 2.1.5 (and inside front cover)
- Substitution – see 2.1.6 (very cool, but be very careful)
- Unix Considerations
  - Grep vs. Egrep vs. Fgrep - if you search for '(B|C)ojack'
    - Grep will search for '(B|C)ojack' as a string
    - Egrep will search for Bojack or Cojack
    - Fgrep will search for '(B|C)ojack' as a string
  - Grep vs. Egrep vs. Fgrep - if you search for '\(B\|C\)ojack'
    - Grep will search for Bojack or Cojack
    - Egrep will search for '(B|C)ojack' as a string
    - Fgrep will search for '\(B\|C\)ojack'

INFORMATION RETRIEVAL EVALUATIONS

- We're building a system that classifies images as either Bojack Horseman or Mr. Peanutbutter
- 100 images, split 50/50 (balanced classes)
- Scenario #1
  - System returns 30 Bojack/15 Mr. Peanutbutter accurately = 45% Accuracy/Precision
  - Majority Class Baseline is 50% Accuracy
  - Accuracy/Precision is fine for balanced classes, worse for unbalanced classes, but ignores full picture of what is retrieved.
    - E.g., if the system only returned 30 Bojacks and 15 Mr. Peanutbutters, accuracy would be 100%, but 55 images went unclassified

INFORMATION RETRIEVAL EVALUATIONS

- Precision ***and*** Recall is a better(?)/ more accepted measure
- Scenario #2
  - System returns 40 Bojacks / 20 Mr. Peanutbutters
    - 35/40 Bojacks were correct = 87.5% precision (TP/TP+FP)
    - 35/50 Bojacks were ***found*** = 70% recall (TP/TP+FN)
    - 12/20 Mr. Peanutbutters were correct = 60.0% precision
    - 12/50 Mr. Peanutbutters were ***found*** = 24% recall
  - System Precision/Recall = 78.3/74.0
  - Often reported as an F-measure, in particular F_1
    - 2 x ((Precision x Recall) / (Precision + Recall))
    - 2 x ((.783 x .740) / (.783 + .740))
    - .760