Midterm Review

Linguistics 384 (Scott Martin)

For the Midterm in class on Wednesday, May 7, 2008

1 Topics to be covered

1. Text & Speech Encoding
2. Searching
3. Spam filtering (document classification)
4. Spelling Correction

2 Format of the exam

You will have the entire 1:48 (1:30-3:18) should you need/want it, but it should be possible to complete it in around one hour.

1. Matching: 10-20 terms (see list below)
2. Calculations: 5-10 questions
    - Binary numbers, ASCII encoding
    - Boolean expressions
    - Regular expressions
    - Precision/Recall
    - Rule-based operations (spam & spelling)
    - Bigram array (positional and non-positional)
    - Confusion matrix
3. Short answer: answer 3–5 out of 5–10
   ⇒ How do the various concepts/technologies work?
    - Types of writing systems
    - ASCII, Unicode
    - ASR & TTS
    - N-grams (spam & spelling)
    - Rule-based spam filters
    - Statistical spam filters
    - Devious spam
    - Types and causes of spelling errors
    - Isolated-word error correction (and its limits)

3 Terms to know

3.1 Text/Speech encoding

- text
- speech
- abjad
- alphabet
- syllabary
- syllabic alphabet
- diacritic
- logographic system
- logogram
- semantic-phonetic compound
- bit
- byte
- Big-Endian
- Little-Endian
- ASCII
- Unicode
- Character encoding
- MIME
- meta-information
- continuous
- discrete
- Hertz
- transcribe
- phonetic alphabet
- coarticulation
- speech flow
- loudness
- intonation
- pitch
- fundamental frequency
- overtone
- spectrogram
- ASR
- TTS
- continuous speech system
- isolated-word system
- acoustic signal processing
- information loss
- irreversible

3.2 Searching

- keyword
- query
- synonym
- boolean expression
- regular expression
- operators
- operator precedence
- escaped character
- counter
- literal strings
- disjunction
- negation
- counters
- wildcard
- linking
- link counting
- formal language
- regular language
- corpus
- meta data
- meta tag
- click-through measurement
- database
- index
- search engine
- relevancy
- precision
- recall
- accuracy
- web crawler
- clustering
- stemming
- capitalization
- ambiguity
- stop words
- web forms
- grep
- (term) weight
- hash table
- part of speech

3.3 Spam filtering/Document classification

- language identification
- document classification
- n-gram
- frequency
- distribution
- spam
- spam filter
- blacklist
- whitelist
- rule-based filtering
- weight
- spam probability
- statistical filtering
- learning
- false positives

3.4 Spell checking

- productivity
- inflection
- tokenization
- detection
- correction
- Spoonerism
- word recognition
- interactive spelling checkers
- automatic spelling correctors
- phonetic errors
- run-on errors
- split errors
- isolated words
- (words in) context
- nonword error detection
- isolated-word error correction
- context-dependent word correction
- dictionary lookup
- dictionary construction
- single-error misspelling
- multi-error misspelling
- array
- positional bigram array
- nonpositional bigram array
- domain-specificity