

Midterm Review

Linguistics 384 (Martin-Lozano)

For the Midterm on Thursday, February 8, 2007

1 Topics to be covered

1. Text & Speech Encoding
2. Searching
3. Document classification (Spam detection)

2 Format of the exam

You will have the entire 1:48 hour should you need/want it, but it should be doable to complete it in around one hour.

1. Matching and Multiple Choice exercises checking your understanding of the concepts and terms introduced in the first part of the quarter (see list below)
2. Calculations:
 - Binary numbers, ASCII encoding
 - Boolean expressions
 - Regular expressions
 - Precision/Recall
 - Rule-based operations
 - Spam probability
 - N-grams

3 Terms to know

3.1 Text/Speech encoding

- | | | |
|-------------|----------------------|-----------------|
| – text | – syllabic alphabet | compound |
| – speech | – diacritic | – bit |
| – abjad | – logographic system | – byte |
| – alphabet | – logogram | – Big-Endian |
| – syllabary | – semantic-phonetic | – 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
- 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