

Final Review

L245

Spring 2017

The final exam will be Monday, May 1, 10:15am–12:15pm, in our usual classroom.

1 Topics to be covered

- Searching (from slide #20 on)
- Document Classification
- Machine Translation (MT)
- Dialogue Systems (slides 1–19 & 31–36 [sections 6.1–6.5.1, 6.6–6.7, & Under the Hood 10])

2 Format of the exam

1. Matching: 10 terms [see below]
2. Multiple choice: 5 questions
3. “Calculations”: 3-7 questions
 - Weblinking & webpage ranking
 - Regular expressions
 - Frequency distributions
 - Evaluation measures (Precision/Recall, Sensitivity/Specificity)
 - Odds ratio calculations for document classification
 - Perceptron weight updating
 - Stylometric analysis
 - Bag of words method
 - Statistical MT
 - Phrase-based translation
 - Lexical relations
 - Dialogue mark-up: adjacency pairs, speech acts
 - ELIZA-like template manipulation
4. Short answer: 5-10 questions
 - Searching in databases vs. on the web vs. in a corpus
 - Structured vs. unstructured vs. semi-structured information

- Uses of regular expressions
- Supervised & Unsupervised learning
- Feature engineering
- Naive Bayes
- Perceptron
- Stylometry
- Author profiling & its consequences
- Authorship attribution (including plagiarism detection)
- Difficulties of (machine) translation
- RBMT: direct transfer systems
- Interlinguas
- SMT: text alignment, bag of words, expectation-maximization algorithm
- Probabilities used in IBM models
- Phrase-based translation
- MT evaluation and uses
- Basic facts about dialogue
- ELIZA/chatbot architecture

3 Terms to know

3.1 Searching

- semi-structured data
- regular expression
- corpus

3.2 Document classification

- document classification
- class
- sentiment analysis
- language identification
- spam filtering
- machine learning
- artificial intelligence
- training data
- test data
- supervised learning
- unsupervised learning
- feature
- model
- prediction
- clustering
- features & feature engineering
- kitchen sink method
- hand-crafted method
- sensitivity (recall)
- specificity
- positive predictive value (precision)
- negative predictive value
- Naive Bayes classifier
- odds ratio
- perceptron classifier
- input/output layers (of the perceptron)
- authorship attribution
- stylometry
- lexical style marker
- function word
- plagiarism
- author profiling
- needle-in-a-haystack
- author verification

3.3 Machine Translation

- translation
- source language
- target language
- machine translation (MT)
- lexical ambiguity
- synonym
- hypernym/hyponym
- light verb
- idiom
- lexical gap
- collocation
- rule-based machine translation (RBMT)
- statistical machine translation (SMT)
- dictionary
- transfer component / comparative grammar
- underlying representation
- interlingua
- machine learning
- text alignment
- word alignment
- bag of words
- phrase-based translation
- intelligibility
- accuracy

3.4 Dialogue Systems

- discourse / dialogue
- utterance
- turn-taking
- adjacency pair
- common ground
- backchannel
- initiating move
- response
- dialog management move
- discourse purpose
- task-based dialogue
- Turing test
- pattern recognition
- chatterbot