The final exam will be Tuesday, May 5, 2:45–4:45pm, in our usual classroom.

1 Topics to be covered

- Searching (from slide 24 on)
- Document Classification
- Machine Translation (MT)
- Dialogue Systems (slides 1–9 & 31–36 [sections 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
   - Regular expressions
   - Finite-state automata
   - Evaluation measures (Precision/Recall, Sensitivity/Specificity)
   - Stylometric analysis
   - Frequency distributions
   - Odds ratio calculations for document classification
   - Bag of words method
   - Statistical MT
   - Phrase-based translation
   - Lexical relations (cf. Mam exercise)
   - ELIZA-like template manipulation

4. Short answer: 5-10 questions
   - Searching in databases vs. on the web vs. in a corpus
   - Uses of linguistic searching
   - Supervised & Unsupervised learning
   - Feature engineering
• Naive Bayes
• Perceptron
• Stylometry
• Deception detection
• Author profiling & its consequences
• Authorship attribution (including plagiarism detection)
• 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
• What makes MT difficult
• Basic facts about dialogue
• ELIZA/chatterbot architecture

3 Terms to know

3.1 Searching
– semi-structured data
– regular expression
– formal language (theory)
– 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
– model
– prediction
– clustering
– features & feature engineering
– kitchen sink method
– hand-crafted method
– sensitivity (recall)
– specificity
– positive predictive value (precision)
– negative predictive value
– Naïve Bayes classifier
– odds ratio
– perceptron classifier
– input/output layers (of the perceptron)
– authorship attribution
– stylometry
– lexical style marker
– function word
– text reuse
– plagiarism
– author profiling
– needle-in-a-haystack
– author verification
– deception detection
3.3 Machine Translation

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

3.4 Dialogue Systems

- discourse / dialogue
- Turing test
- pattern recognition
- chatterbot