***Agenda – Week 10 (December 7, 2020)***

* ***Welcome***
* ***Information Retrieval (IR)***
  + ***Finding material (usually documents) both structured and unstructured in nature that satisfies an information need***
  + ***IR Systems*** 
    - ***Process documents and identify terms to be indexed***
      * ***Stemming applied [in order to widen the search of the applicable terms] and Stop Words are removed***
    - ***Compute the weights of terms (keywords) based on a model definition***
      * ***Example Models***
        + ***Term Frequency (TF) = # of times the term occurs in the document [freqtd] divided by the length of the document [lengthd] which normalizes the result; TF = freqtd / lenghtd***
        + ***Inverted Document Frequency (IDF) = calculation based on the number of documents in a collection [N] and the number of documents with the term [nt]; IDF = log(N/nt)***
    - ***Build an index of documents and weights***
    - ***Calculate a relevance score for the document based on its weight***
    - ***Order the documents returned in the order of their relevance score***
    - ***Evaluate the relevance score:***
      * ***Precision = a / (a + b), where a=TP, b=FP, c=FN and d=TN***
      * ***Recall = a / (a + c), where a=TP, b=FP, c=FN and d=TN***
      * ***Mean Average Precision***
    - ***Improving Results***
      * ***Query Expansion – adding semantically similar words the query***
      * ***Relevance Feedback – query the top “n” documents determined to be relevant with additional terms***
  + ***Web Search – form of IR where the collection of documents are located on the Web***
    - ***Vocabulary Gap – words in the document might not match the query***
    - ***Need to determine the documents that are most relevant (i.e. ranking)***
    - ***Setting Robustness & Politeness thresholds important for performance (i.e. retrieving results and not overwhelming the web server)***
    - ***Can build your own web crawler or spider that adhere to or are constrained by the threshold settings***
    - ***May add keywords to improve search results (e.g., metadata, anchor text, and weights for tokens appearing in the header tags***
    - ***Improving Results***
      * ***Scoring algorithms for webpages (e.g., PageRank & HITS)***
      * ***Popularity – current / trending topics***
      * ***Click-Through Rates – using statistics on pages users access***
      * ***Context – tracking user interests, location , etc.***
      * ***Machine Learning – use ranked relevance (achieved by above) to improve rankings***
    - ***Discounted Cumulative Gain (DCG) evaluation measure***
      * ***Measures relevance at each ranked position***
      * ***Penalizes for incorrectly ranked documents***
      * ***A/B Testing***
        + ***Create to versions of the web search algorithm***

***Group A – Regular Search***

***Group B – Experimental Search***

* + - * + ***Evaluate & Adjust Rankings***
    - ***NLP increasingly being used in web search queries*** 
      * ***Stemming, stop words, named entity recognition, etc. are used in the search algorithms***
* ***Question Answering (QA)***
  + ***Similar to IR in that the user is retrieving and analyzing documents. The difference being with QA, the user only wants the answer to the question found in a specific document*** 
    - ***Parsing / Semantic Analysis may be applied to get answers***
  + ***Build question ontology (i.e. properties)***
  + ***Look for “named entities” (e.g., PER, LOC, etc.) to answer questions***
  + ***Specialty Answer Processing*** 
    - ***Define patterns***
      * + ***Often use hand-built regular expression w/defined patterns***
  + ***Improving Results***
    - ***Answer Ranking***
    - ***Confidence Levels***
  + ***Watson Question Answering System***
    - ***Deep QA Project*** 
      * ***Used to play Jeopardy***
    - ***Need to store a significant amount of data (pre-computed knowledge)***
    - ***Needed to be flexible to the type of Jeopardy-style question***
    - ***Excellent example to NLP architecture to accomplish the analysis and answer response***
    - ***Built in decision and confidence criteria used to determine whether Watson would buzz in and amount of wager on Daily Doubles and Final Jeopardy***
* ***Conversational Agents***
  + ***Task-Oriented Dialogue Agents***
    - * ***Digital Assistants – Alexa, Google Assistant, Siri, etc.*** 
        + ***Frame-based***
        + ***Preprocessing of the conversation***
        + ***Fill into predetermined frames***
        + ***Determine response / including error response***
    - ***VoiceXML – digital document standard for specifying interactive media and voice dialogs between humans and computers. It is used for developing audio and voice response applications, such as banking systems and automated customer service portals.***
  + ***Chatbots***
    - ***Assessment: Turing Test – passing for a human***
    - ***ELIZA – early rule-based chatbot***
* ***Questions***
* ***Lab Session 10.4 – Setting Up Different Classification Tasks For Use on Final Project***
* ***Open Discussion***
* ***Assignments Due:***
  + ***NLP Application Investigation – Week 10 (12/7) & Week 11 (12/14)***
* ***Looking Ahead:***
  + ***Final Project: Classification of Text – Due 12/16***
* ***Next Class: NLP Application Investigation – Student Presentations***
* ***Closing Remarks***

***Assignments & Due Dates:***

* ***Homework Assignment 1: Corpus Statistics and Mutual Information – Week 4 – Due 10/26***
* ***Homework Assignment 2: Regular Expressions – Week 6 – Due 11/9***
* ***Homework Assignment 3: Context-Free Grammars – Week 8 – Due 11/23***
* ***NLP Application Investigation – Student Presentations – Week 10 (12/7) & Week 11 (12/14)***
* ***Final Project: Classification of Text – Due 12/16***

***## Assignments can be completed ahead of the published due dates.***