605.744: Information Retrieval
Problem Set (Module 11)

Sabbir Ahmed

November 21, 2022

1. (30%) **Question Typology.** IBM's Watson team developed over a hundred ways to answer a Jeopardy! question. A key part of question answering is determining what kind of question is being asked. Mark each question below with one of the following question categories: QUANTITY; WHERE; WHEN; WHO; WHAT-SUBTYPE; OTHER. Finally, write a regular expressions (or create a pattern like a regular expression) that correctly identifies all of the QUANTITY questions below but does not mis-identify any of the other questions as QUANTITY.

   (a) At what temperature does water freeze?
   Answer: QUANTITY

   (b) How tall is the Eiffel tower in Lisbon?
   Answer: QUANTITY

   (c) Who killed Bobby Kennedy?
   Answer: WHO

   (d) How far is the Earth from the Sun?
   Answer: QUANTITY

   (e) What flying mammal navigates using echolocation?
   Answer: WHO

   (f) Where did Elena Kagan attend law school?
   Answer: WHERE

   (g) What US federal agency is responsible for collecting tariffs?
   Answer: WHAT-SUBTYPE

   (h) Who invented Post-It notes?
   Answer: WHO

   (i) How do you change the oil in a Ford Explorer?
   Answer: HOW

   (j) When did Idaho become a state?
   Answer: WHEN

   (k) Which national monument did Dr. Martin Luther King Jr. give a famous speech at?
   Answer: WHERE

(l) How tall is the belfry in Bruges?
Answer: QUANTITY

(m) What city was Angela Merkel born in?
Answer: WHERE

(n) How many feet are there in a nautical mile?
Answer: QUANTITY

(o) What is the state flower of Maryland?
Answer: WHAT-SUBTYPE

Answer: The regular expression would be: `(How\s(?!do))|(What\stemperature)`.

2. (25%) **Named Entity Recognition (NER).** A NER tool tries to identify named entities in text (i.e., persons, organizations, locations). Examine the online Stanford NER tool available at: https://corenlp.run. Enter a variety of text and examine the results.

   (a) Give an example of an input sentence and error that you think the tool should not make.
   Answer: *I can't wait to try cooking a turkey this Thanksgiving!*

   (b) What types of errors are you able to discover?
   Answer: The named entity *Thanksgiving* was properly recognized as a `DATE` entity, but *turkey* gets improperly categorized as a `COUNTRY` entity.

   (c) Now suppose that you had a perfect named-entity recognizer (i.e., one that makes no mistakes). Argue briefly and clearly whether or not this capability could be used to effectively enhance ad hoc text retrieval accuracy (i.e., as measured by say average precision). Explain your reasoning and give examples if helpful.
   Answer: If a user queries *turkey* in a retrieval system that emphasized its perfect named-entity recognition, documents that recognize the term as only a named entity will be retrieved. i.e., documents containing topics on the named entity version of the word (the country) would be retrieved while ignoring documents about the species of bird.

3. (15%) **Retrieving with Good Sense.** Read Mark Sanderson's paper "Retrieving with Good Sense" (this is an assigned reading). In a few sentences briefly describe Sanderson's kalishnikov/banana experiment. You should explain what the goal of the experiment was and what was learned.

   Answer: Sanderson attempted to artificially add ambiguity to a corpus by joining random words together to create pseudowords. The constituent words are then replaced by the pseudoword, i.e. all occurrences of "kalishnikov" and "banana" in the corpus get replaced by "kalishnikov/banana". Sanderson used this approach to add n-sized pseudowords (pseudowords made of n distinct words) and determine the effectiveness of an IR system. It was found that adding pseudowords did not reduce effectiveness as much as might have been expected.

4. (15%) **Lexical Semantic Relations.** Make up your own examples for the following lexical semantic relations. Example: "Two words that share an antonymy relation" - excited / calm

   - Two words that share a hypernym / hyponym relation:
   Answer: color / red

- Two words that share a demonym relation:
  Answer: Spain / Spanish
- Two words that share a synonymy relation:
  Answer: copy / duplicate
- Two words that share a meronymy relation:
  Answer: book / library
- Two words that share a troponymy relation:
  Answer: laugh / giggle

5. (15%) **Using WordNet.** Explore the on-line version of WordNet, which can be found at http://wordnetweb.princeton.edu/perl/webwn. Lookup the detailed entries for these words: alphabet, delta, oracle, yeti.

   Given what you observe by looking up these words, what conclusions can you make about using dictionary-based word-sense disambiguation (e.g., using a resource like WordNet) to try and improve text retrieval performance?

   Answer: Using semantic relations databases like WordNet can drastically improve text retrieval performance. Words such as "yeti" only had a single definition and can therefore be disambiguated very easily without the need for additional word-sensing. The other words have multiple definitions and would require WordNet to resolve the context. Using word-sensing instead of the frequency of a version of the word would yield more accurate text-retrieval results.