Class: Question answering (chapter 23)

December 2, 2010

Admistrivia

• Status reports?

Who is John Galt?

• Goal: Automatic question answering
• Incoperates all that we have discussed so far
• General approach:
  – Find documents that are related to the question
  – extract relevant information from documents
  – generate answer (often via template) “John Galt was born on <birthdate> in <birthlocation>. He is famous for <significationfactoid>....”
• We will take each of these in turn

Information retrieval: I.e. Google

What we do

• If we are building such a system, we will send queries to google to get back pages

• They have a nice API to support this

• They are willing to accept several 1000 queries a day before they want money

What they do

• How do you do this by hand? (or for your new internet startup company)

• Easiest approach is the vector method

Vectors

‘Who is John Galt?’

• represent as a vector with all zeros and 4 ones in it

• represent each document as a vector of counts

• How close are these two vectors?
cosine

How close are two vectors?

- Longer articles and shorter articles should be normalized to same length
- This leads to a cosine metric.

\[ ||X - Y||^2 = ||X||^2 + ||Y||^2 - 2X^tY \]

So if we have normalized \( ||Y|| = 1 \) and we use the same \( X \) over and over again, minimizing \( ||X - Y/\sqrt{||Y||}|| \) is the same as maximizing \( X^tY/\sqrt{||Y||} || \).

IDF

- But, “who” and “Is” are uninteresting, “John” isn’t that useful either, but “Galt” is now a winner.
  - \( n(“who”) = 4.4 \) B
  - \( n(“is”) = 11 \) B
  - \( n(“john”) = 1 \) B
  - \( n(“galt”) = .022 \) B
- But the spread is even richer. The “who”’s are spread all over the web, whereas the “Galt” are highly concentrated. So a article which has one “Galt”, probably has 10 “Galt”’s.
• Idea of document frequency. Rare document words are a fingerprint for a document.

• Called IDF. Typically log’ed for stability

Putting it together

• TF/IDF: Term frequency / inverse document frequency

• Then use cosine

• Grab top hits

Better top hits

Cool idea: Spread out top hits to cover different topics / pages. David and Ben discussed this a few weeks ago in MLunch. Instead of sampling based on the TF/IDF, sample based on the determinant of closeness of top hit documents.

Next semester, those who want to continue in this direction, should try to attend both MLunch and CLunch. Free food and good talks (at least some of the time) and good connections.

Factoid questions

Cool idea

• Search for “Gandhi 1869” or “Gandhi 1869 birth”
• You will find 100’s of different ways of describing the fact that Gandhi was born in 1869.

• These can now be used as patterns for finding other birthdates

• Deep CCA:

  left vector has

  – Gandhi 1869
  – Lincoln 1809
  – JFK 1917
  – ...

  Right vectors:

  – born on
  – <person> (#### - #######)
  – JFK 1917
  – In May of 1962, Marilyn Monroe helped to celebrate President John F. Kennedy’s 45th birthday at a spectacular party in Madison Square Garden.
  – ...

• CCA between concepts and syntax
Template answers

We then want to generate answers like, “<person> was born in <year>”. If we can summarize information well enough, we can force it to generate sentences of this nature. Currently, this doesn’t work so well. Alternatively, we can view it as a “fill in the blank.” This now is at least interesting statistics. Converts the problem into a “cloze” question.

Cloze data

on sobolev, we have lots of “the” vs “her” fill in the blanks. These are very simple close questions. But we have to be able to solve these before we solve the harder stuff.