Conversation Clustering and Search

Greg Tourville

Background

- Search is obviously important
- Better search quality
- Existing methods:
  - Bag of words
  - Vector space model
- Why?
- Limitations

Websites vs. Conversations

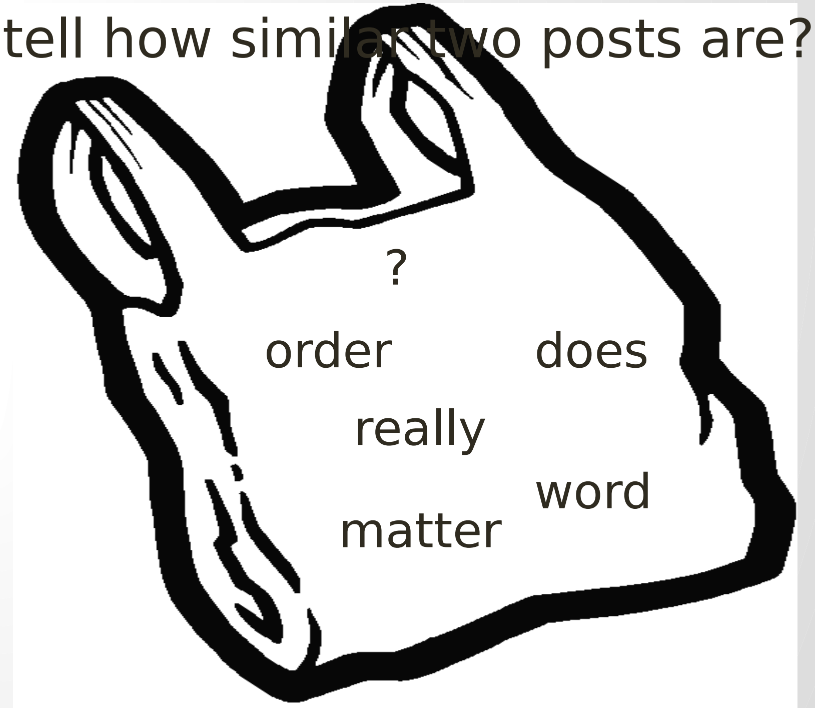
- The internet talks like an encyclopedia
- People do not talk like encyclopedias
- Websites have SEO
- If you’re optimizing your conversations for Skynet that’s weird

What is conversation?

- 30,000 comments from the reddit front page
- Each individual comment is really its own document
- Examples:
- "All dictatorships rely on a vast network of people that owe their wealth/well-being to the existence of the regime. That is one of the reasons why dictatorships are always full of corruption.“
- “Carrots are not naturally orange. The wild plant was purple or yellow, the orange variety was selectively bred by the Dutch around the 17th century.”
- You learn something new everyday

Ideas we already have

- Hierarchical Clustering
  - It sounds fancy because it is
  - Probably really hard to do with lots of data
  - How do you tell how similar two posts are?
- Bag of words



?

order does

really

word

matter

Reading deeper

- "Recently the Graduate Employees Organization at the University of Illinois went on strike to protect tuition waivers for graduate teaching assistants. Considering UI uses graduates to teach a large % of their classes, I'd say it's all fair. Especially considering the cost of hiring a professor for those classes.“
- graduate, employees, organization, university, illinois, strike, tuition, graduate, teaching, UI, graduates, teach, classes, cost, hiring, professor, classes
- How can we get this post to show up under the search "illinois college education union" ?

What are we really talking about

- We develop jargon for different topics
- Related jargon appears together in conversation
- Statistics lets us know more with less
- The internet’s thesaurus
- Stemming
  - This works pretty well for this actually

New ideas

- If we have two many documents to cluster, why not cluster words instead?
- We can assume new words come up less than new comments
- Words are added as they are used, telling us more about them
- Ouroborous

© Saki 04 2005

Math and word games

- Let A and B be words. We can define A’s relation to B as:
- $P(B|A) \times idf(B)$
- $idf(B) = 1/\log(|B|)$

Things machines can do

- Keep tallies of how all words are used together
- Oxford says there are 200,000 words, so that would be 160 GB as an array of ints
- But storage is cheap
- And JavaScript makes writing code easy, especially with associative arrays

What this gives us

- We end up with a dynamic thesaurus
- It includes made up words
- Somehow JavaScript can process more than the bible worth of random text in a minute
- Ways to use this:
  - Feed it into the google net
  - Or go back to those fancy clustering algorithms
  - Easier for words, then plug and chug comments through them
- Depends on the application
  - Are you looking at everything ever said?
  - Or do you just want to find people to talk to

Magic?

- NLP is hard it turns out
- We’re getting closer, but the answer is probably duct tape
- Easiest way to get better results:
  - Stop list
  - Changing 1/log(B) to 1/log(B+4)
  - Using sqrt() instead makes the results suck
  - Hanging out with George Zipf

Being realistic

- Let’s look at it running and see how it kinda works:
- fructose: 17 study; sweetener; obesity; rats; syrup; glucose; sucrose; sugar; hfcs;
- flown: 4 going; stall; landing; hours; followed; idea; plane; flying; fly;
- landing: 4 x; x; x; x; nervous; stall; look; likely; plane;
- cops: 103 road; lights; corner; officer; guy; police; couple; driving; pulled;
- unable: 29 unemployed; used; unemployment; whining; unless; underground; unreasonable; underlying; uneducated;
- elementary: 35 gym; turtle; library; jr; junior; teacher; middle; high; school;

Future work

- Lots and lots of tweaking
- Maybe some genetic algorithms and human training
- N-grams where n > 1
- Looking through this data is a lot of fun
- ?????
  - Topic, Tone, Meaning, Aboutness, Grammar, Thought, Nihilism
- The Singularity

Questions?