HAYSTACK
The Search Relevance Conference

Semantic Knowledge Graphs for Search

"Populating and leveraging semantic knowledge graphs to supercharge search"

Chris Morley
OpenSource Connections

April 2023

Experimental content ahead…

Where we're going, we need lots of roads, actually.

This talk is very new and I hope that I can give it many more times and improve it in the future, so any and all of your feedback is super valuable. There is lots of theory in here and a bunch of ideas, and specific examples. Actual implementation of all of these ideas is still evolving for me and some things in here may be things you have encountered but the person next to you hasn't, and some other things it might be the other way around. I have a quick POC working but there is much more in the future on this than in the can, so to speak! How "it is done" is not as important as how you might want to suit these ideas to fit your own needs. We can't make things too generic without trying it a few times in some specific cases.

Wormholes

Buckle up

There is just too much to say, really.

About Me

Chris Morley
Search Relevance Engineer
OpenSource Connections
Professional software developer since 2000
Tinkering w/computers almost constantly since 1988! (Not proud.)
From Wellesley, near Boston, Massachusetts, which is in New England (N.E. U.S.A.)
(Known as “Skipper” or “Skip” when I am on Cape Cod)
Wife Melisa a licensed optician.
Two children Xander (20) and Phoebe (18)
cmorley@opensourceconnections.com
@depahelix

Chris Morley

OpenSource Connections

We are your Haystack hosts!

We empower the world's search teams to be able to operate their search engines effectively and we want to help you understand how to make your search results be super awesome for the people who use your search system. The exact KPIs that determine what your success is going to look like is going to be specific to your organization and your context, and we like to help you clarify that vision and help you get moving towards your ambitious goals!

This is a rapid-fire all-over-the-map, theory-heavy synthesis survey that's not really trying to be about "one thing"…There's lots of things to touch on. ☺

If you're like me, and loved Trey Grainger's Keynote, I hope you will also like this talk very much!

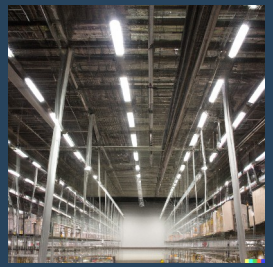
Knowledge Graphs, Character and Term Sequences
Ontologies, Taxonomies, Multiclass Classification
E-Commerce Categorization
Semantics and Linguistics
Cognitive Science and Computer Science and a bit of Philosophy
Named Entity Recognition (but perhaps a bit different?)
Multiword Synonyms, Tries

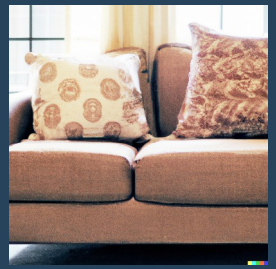
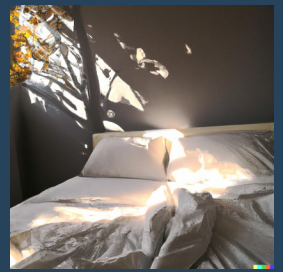
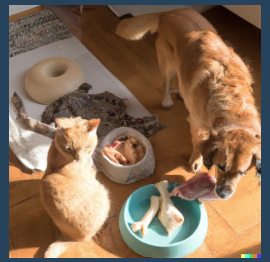
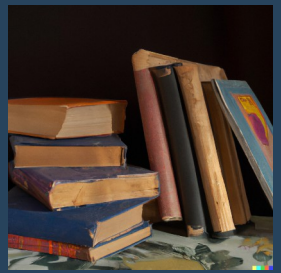
Pointing out areas of opportunity in Search and general suggestions!

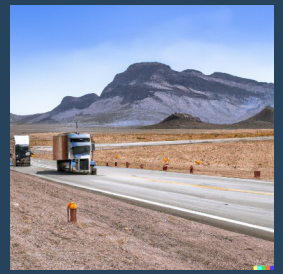
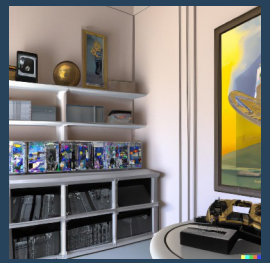
The deck is too big, so you some stuff will just zoom by.
Don't worry about it.

E-Commerce search background 2011-Present

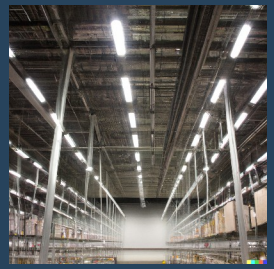
Some background of mine about why I am so interested in graphs as they pertain to search engines:



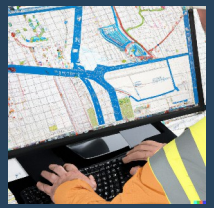





Lightbulbs

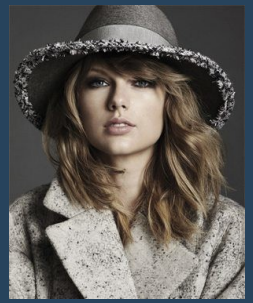
- tons of variety!
- first exposure to search because SQL was way too slow!
- search on a map for businesses by criteria
- literal traveling salespeople planning routes
- incentive programs – business hours + quick survey ==> $$$ for you!

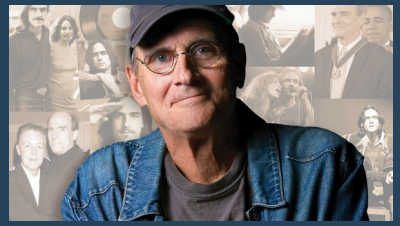


Digital Books, Music & Movies 2012-14

- searching across different types of things that have different origination schemas, merge under **one type of common object**
- artist names: sometimes listed/searched last then first or first then last
- got me thinking about phrases and term sequences
- search to power browse

Taylor Swift

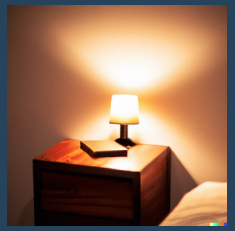




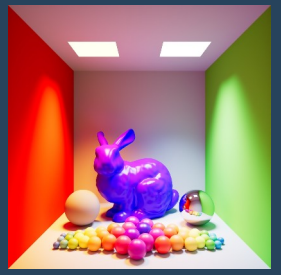
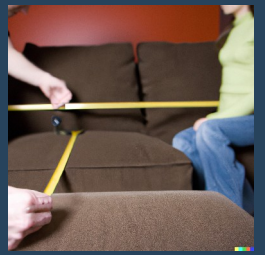
James Taylor

Furniture + Home Queries 2014-16

- search + browse
- big site, many exemplary problems

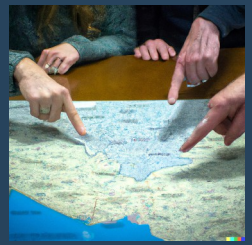


Transportation/Logistics Search 2017

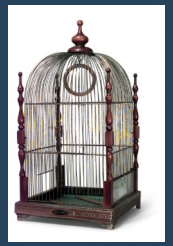
- short and long-haul shipping (big rigs, 18 wheelers)
- ZIP to ZIP, zone to zone, ZIP to zone, estimate accurate quotes
- innumerable potential combinations
- short or a long, easy or tough drive
- polygonal geometries
- find past business contracts that were similar (“graph-y” problem)
- my data visualizations helped w/ “Where should new (Amazon) fulfillment centers go?”

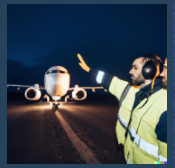


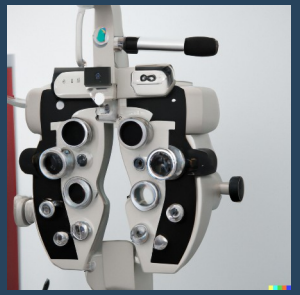




Pets 2018-20

- “Halloween”
- rabbit food
- redirect rules
- expansion / relaxation
- products, items, categories
- other things:
  - recency vs relevance
  - implicit judgements
  - a/b experiments

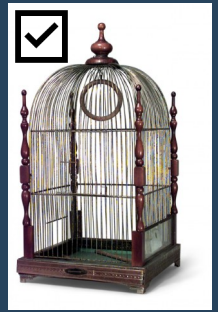






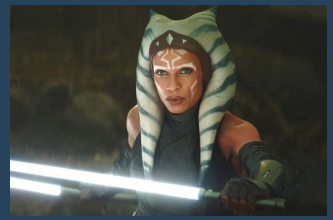
Yrllensdan != Dog Food : aka *OOV*

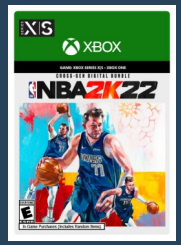
- hypothetical reminds me of real issue
- Walmart exclusive brand: crates, shelving and cages
- suppose you get "**yrllensdan**" (a brand you don't and won't carry),
  - what do you do?
- "Hey there! Say, how would you like some nice *dog food*?"
- User: Huh??



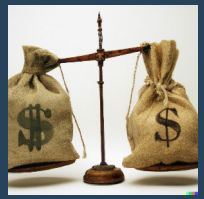
Video Games 2021-22

- nba2k22 vs nba 2022 (Grandma vs. nomenclature?)
- ahsoka vs ashoka (common misspelling)
- Wolverine is in the X-Men!





FST autocomplete
a **weighted** trie graph



So why semantic knowledge graphs?

- **search engines** are *great* at what they are good at, but
  - for certain types of activities search engines are **weak**:
    - **joins**, linked data, parent child, nesting – struggles
    - **shingles**, way too many – n-grams out of control, performance issues
    - **out-of-vocabulary,** unless you have hidden indexes?
    - **query intent / q parsing / q understanding / q segmentation** – custom plugins?
    - **peek @ stats?**, lacks stats, metadata to predict performance from super recent performance : maybe w/ custom wrapper services?
    - **naïve whitespace tokenization** "buzzsaws" meaning
    - **synonyms are tough**, *most especially multi-word synonyms*

Graph DBs

- they **shine** at dealing highly interconnected data, hierarchical stuff (strong where search engines are weak)
- many Graph DBs have Lucene under the hood, but never fear, they are definitely not here to replace your Solr, Elasticsearch or OpenSearch
- to be used in addition to, **combined with**, not instead of
- There can be graph features in search engines as well, but they aren't typically fully featured like a Neo4J or similar engines (no offense!)

Graph Databases, Generally

What is a graph database about anyway?

Two First Class Primitives:

NODES: aka vertices, entities, objects, rows, locations, concepts, things, waypoints, circles, spheres, planets, points in space, (perhaps: documents? mappable Very loosely to vector embeddings?)
EDGES: aka relationships, links, lines, joins, connections, shortcuts, arrows, paths, roads, ways, routes, branches, wormholes, “is a”, “has a”, …and any other ways you can imagine

Many Graph Use Cases

- resolving user permissions from group membership and group permissions – (Please ask Kevin Watters about this!)
- master data management – tracking data from multiple sources and merging it to ensure uniformity, accuracy, stewardship, semantic consistency and accountability of official shared data
- anti-fraud: example - calculating that multiple shady businesses and/or aliases are originating from the same physical address
- social networks – self joined objects – person “knows” person
- *taxonomies, ontologies and semantics, crystalized/materialized clusters, potential shortcuts for vector similarities, KNN, ANN, faster?, more understandable, totally explainable! Do stuff to stuff.*

Graph DB we'll use: Neo4J

PROS:

- very fast, in-memory DB
- very flexible (think "schema-less")
- good for highly linked interconnected kinds of data
- good for hierarchical, parent/child kinds of things, self-joined things
- query language Cypher is *powerful*

CONS

- subject to memory limits
- expensive to load/curate/maintain (*changing nowadays with advances?*)
- query language Cypher takes some getting used to

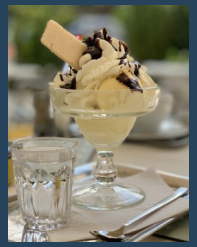
What problems do we want to solve?

- Help to resolve SEMANTIC issues at index time or even prior to that (schema design time?)

- Quickly (near real time) resolve SEMANTIC issues at query time before we issue a search to our search engine

- COMPOUND WORDS
- CONCEPTUAL SYNONYMS, CONCEPTUAL TOKENS
- FIELD MATCHING
- CRYSTALIZED CLUSTERS / CLASSIFIERS STAMPED

Compound Nouns and Noun Phrases

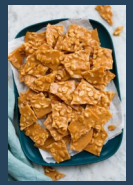


- words made up of multiple words can be separated by space, hyphen or nothing, or a combination of those:

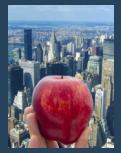
ice cream…sundae?
- root beer…float
- toothbrush…case/holder

football…league/shirt/player
- button-down shirt…sale/clearance

- July 4th / Fourth of July
- Independence Day



low calorie vs sugar-free
peanut brittle vs peanut bark vs peanut crunch
vs peanut butter … and jelly?



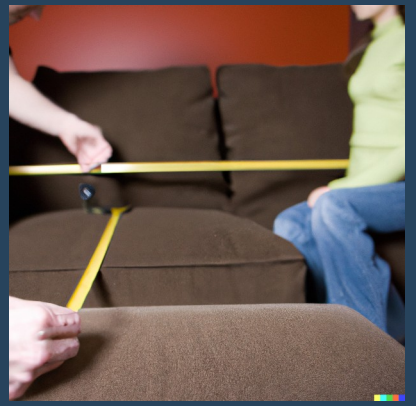
New York City, NYC, big apple
the city so nice they named it twice

Compounding and Decompounding

- German decompounding – breaking up long words into pieces
  - Are we talking about something like a “compounding compounder” for English? Compound noun and noun phrase “aware” tokenization by way of a dictionary, trie, Aho-Corasick?
- some Asian languages do not use space at all
- get single token *only when appropriate*
  - not everywhere, let’s not go overboard with n-gram shingles
  - vectors / embeddings accidentally work, but you can only do certain things with them, you cannot make logical, lexical leaps

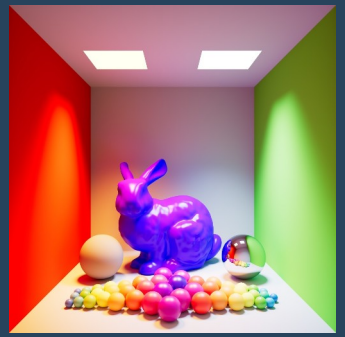
Size / Dimensions

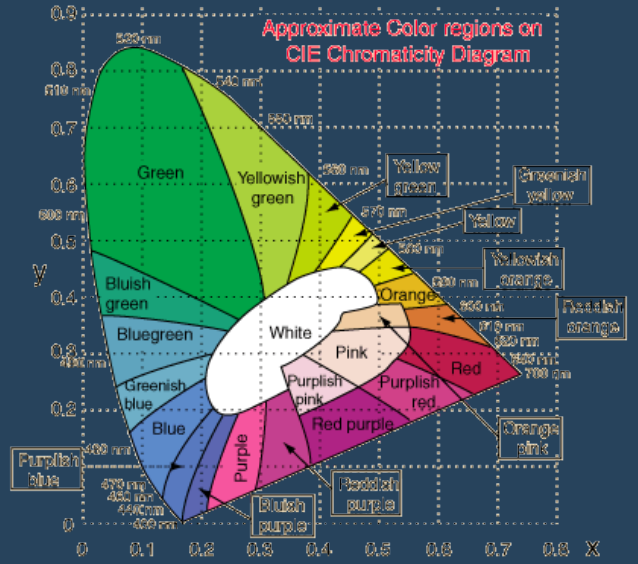
- a large cat vs small dog, for example?
- L W H D!?
- Units!!
- Canonicalize for contextual items?



Colors, Colors and more Colors

- innumerable nicknames
- certain canonical hex colors as “magnets”
- paint vs. products
- Vast “out of vocabulary” potential





Recap

- compound nouns and noun phrases, noun chunks
- conceptual synonyms
- matching types of strings to types of fields
- sizes / dimensions
- simplify color space!

Mixed up letters: Heart vs Earth

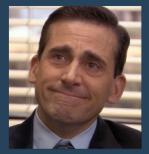
- they have the same letters but do not mean the same thing
- many people can read words if the first letter and last letter are fixed but the letters in the middle are scrambled, especially in sentences:
  - my hraet baets lulody wehn i stnad on teh etarh.
  - but if you mix up **ALL** the letters, that's a classic anagram puzzle
- emoji movie descriptions – too short and simple, mix up, sure
- some sequences still work when you mix them up, others, not really
- jelly and peanut butter sandwich just doesn't have the same ring to it

HAYSTACK
The Search Relevance Conference

Mixed Up Words

- if a mischievous leprechaun or gremlin mixed up the order of words in each sentence of to play a prank on you, but *what if all major phrases were kept intact?*

- whitespace tokenization destroys meaning of compound nouns
  - ice cream never means The Cream Ice Shelf
  - peanut butter is its own thing apart from peanut and/or butter
  - root beer float never means a beer themed pool float
  - sometimes some words just go better together



Word order ALSO matters: "Royal Order of Adjectives"

- People learning English as a second language try to memorize
- Native speakers know this already automagically without realizing it
  - it gets sort of hard wired, in-grained
- Incidentally it goes: opinion-size-age-shape-color-origin-material-purpose

- Example:
  - "We are going to have a **big Italian** meal." => OK, sounds good!
  - "We are going to have an **Italian big** meal." => Wait, what?

  - THIS IS JUST ONE EXAMPLE. Linguistics and simple thought experiments could yield countless other types of examples that are similar but different.

Noun chunks

- a stick of gum
- a pack of gum
- big red ball
- the leafy green tree

- noun chunks is a terminology that spaCy uses

- vectors / embeddings good by accident because of wide char spans,
  - “mutual information” leaking in via training process
  - things not truly understood, but exhibiting mimicry of what a system could do were it to be understanding

Scary OOV

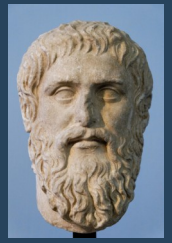
- ***caiman*** (**kay**·muhn)
  - it's basically an **alligator**
  - for all intents and purposes,
    - we do not care about the distinction!
- ***gharial***
  - (**geh**·ree·uhl)
  - basically a **crocodile**

- *later vs. after a while*

~~Alligator~~
~~Croc~~
~~Crocodile~~
~~Caiman~~
~~Gharial~~
~~Gator~~

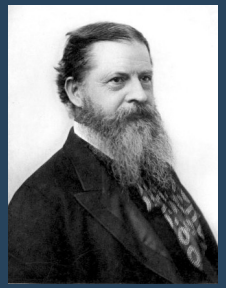
Sweet Everglades Shirt ~~$17.99~~ $17.99

"Semeiotics" in Three Domains



Plato:
"Platonic Ideal"
aka "OBJECT"

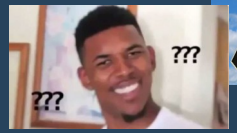
- **Charles Sanders Peirce (1839-1914)**



- Object Domain
- Sign Domain
- **Interpretant** Domain

Somebody says out loud or writes something down, making some "SIGN"

Interpretants often left guessing what the heck was ever even meant…



Domain specific language/jargon

- overloading
- misunderstandings + ambiguity
- TLAs, code words, closed doors?
- Wait, what are we even talking about again? Different departments!

..or at least getting a slightly different idea. "Playing Telephone"

Speaking of which…

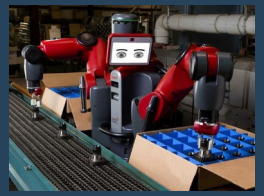
- I hope that at least some of what I'm saying is coming through clearly.

If not, oh well…
That's to be expected.
It's kind of the default setting
of the universe.
Nobody *really* truly and fully
understands anybody else, really, ever.
The only one who will ever know
what it is *exactly* like to be you is you,
obviously.

Conceptual synonyms

- Sign: “he who shall not be named” => Object: Voldemort
- Sign: “the big apple” => Object: New York City
- Sign: “the city so nice they named it twice” =>
  Also object: New York City
- Sign: “fourth of july tablecloth” => Object: tablecloth that is red, white and blue or that features the American flag on it, or fireworks, or the Statue of Liberty, is patriotic, or is otherwise in the “Independence Day” category,,, lots of criteria… lots of weird amorphous boundaries?

Old tricks – commonly accepted as OK



- Stemming & Lemmatization

- These work on the principle that words with common roots can sort of mean the same thing, in a way, but these only barely scratch the surface of synonymous meaning, by sheer luck:
  - oh you said “run”? I’m going to match “ran” also.
    - Our machines: Look at me!
      - I appear to know something useful!

- The opportunity is to do the same sort of thing but for **way more** things of meaning, things that matter: even ideas w/arbitrary complexity!

More tricks – dirty tricks “unacceptable!!”

- Labelling docs with previous top queries that should match the doc?
- Keyword stuffing – that’s just uncalled for!
- Also work on the principle of “luck” / “bidding to win”
- These tricks do not scale and are bad for business, unreproducible
- Very much SEO spam related, driven by wrong incentives

Product Codes vs Product Types – likely field matching via pre-lookup

- Very generally, let's say that we have some product codes:
  - MK-1846572 and XD-9390234
- Let's suppose the MK one refers to a chair, and the XD refers to a table and we get reasonable (not trick question) searches like "chair MK-1846572", "table XD-9390234"
- Why would we go looking for "table" or "chair" in the product code field when we could have recognized/matched those strings as not looking like a product code? Or, why would we look for XD-9390234 or MK-1846572 in the title field or product type field when logic dictates that that will never work (?)
- Is it just laziness that causes us to go and waste of cycles that could be avoided with some pretty simple logic in our pipeline? Not even sophisticated logic is necessary, just some forethought and some logical planning for this could work – example: Redis lookups?

Term search vs Field search? Best vs. Cross??

- Instead of choosing to have our cake or eat it, maybe we can have both????
- Search for terms only inside a specific set of fields that are **most likely** to contain such information!
- Better than a catch all, can do specific boosts based on confidence, statistics known in graph

So, uh, how about…

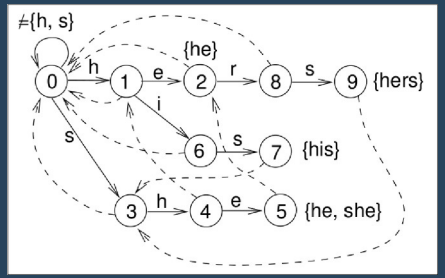
Conceptual emblems?!

- Do not search for this or that or that or that or that or that….it's way too many combinations! (query expansion, phrase recombinatorics, synonym graph, escalates *quickly* and accidentally flies off the handle)
- **Let's just go and find the things we are actually talking about**, (hopefully tokenized as such), on documents that we have already labelled with all the things (again, use good, conceptual tokens / emblems / stamps, domain specific, domain meaningful!)
- Use traditional, classical, lexical tokens for new life as embedding cluster or miniature categorical signposts.
- Yes! this is certified authentic and good for this specific thing, e.g. Season, Holiday, yada yada yada, "what_have_you"…. Redirect all "random" other input to nearest signposts

“All roads lead to Rome.”

- “If the shoe fits, wear it.”
- Gravity, electricity and magnetism exist. Simulate, emulate. Mimic.

Aho-Corasick



- parsing algorithm from the 1970s
- identify all the words in a piece of text that match any of the words in a given dictionary, in a single pass, by building a trie graph data structure from the given dictionary, in advance, before parsing begins
- we need this same thing at index and query time, but at the word level instead of at the character level (letter level might also be useful for guessing which real word a misspelled word would best match)
- Wait - Tries inside our Knowledge Graph? => UM, YES PLEASE!

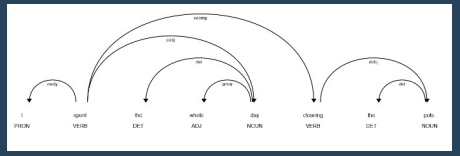
NLP Techniques 1

- Leveraging existing dictionaries
  - common words and terminologies
  - domain words and terminologies
- Leverage existing ontologies
- Ingest competition's website, ingest general domain docs/pages
  - Fight the Out-of-Vocabulary problem by just knowing about more stuff in the world, outside the words in our little world (our specific ecommerce catalog, our specific, limited corpus)

NLP Techniques 2

- Looking at word *transition* statistics, *word chains*
- Linking nearby words together, when appropriate
- Prune after a while. Ignore or forget granular statistics about things that don't happen much anyway
- Count word occurrences generally before the search engine does it, specifically

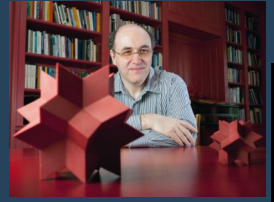
NLP Techniques 3

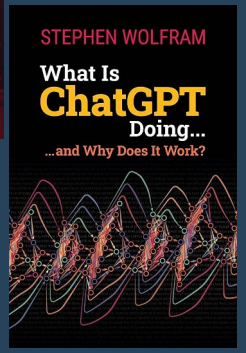


- Dependency Parsing + Dependency Graph information. Sometimes words that are far apart in a sentence actually are more connected to each other than the words right next to them. Pronouns refer to “something else”.

- Precompute stats before indexing / reindexing

- Use at query time too - guess better things to match what the person is actually talking about, trying as much as possible to capture rather than ride roughshod over their meaning





NLP Techniques 4

- When Leveraging LLMs…
  - ChatGPT, AutoGPT, etc.
  - *sub word* tokens, not multiple word concept tokens, so ideas here are different, combines well with these new “powers”
  - Slow can mean “expensive”
  - Good at dealing with the gray maybe/probably: not T/F for sure
  - Non-deterministic - 80% parameter – follow this “interesting” path
  - Some “facts” are just made up out of whole cloth aka hallucinated!
  - Curate, correct and store in semantic knowledge graph
    - Truth: Facts we think we “know” to be true with enough certainty (in real life, people also think it’s the case) to let the system act as if they truly are true. Not things that machines just decided 10 milliseconds ago probabilistically ought to be the case.
  - Not saying don’t use them at all. Just saying be careful: they are exciting and frightening all at once.

NLP Techniques 5

- Don’t ignore small, common words
- These are often “function words” and they might actually really help give you and your parsers and analyzers extremely valuable hints about the surrounding words, just like they do for early readers
- Pay very close attention to proximity and shared context, but also, keep an open mind and remember, you cannot “actually” ever get a tempest to fit in a teapot much less a teacup.
- It is always going to be a little bit wrong.
- It will never be perfect.
- It will never be done.
- Continually strive to be “less and less wrong”.

Function Words

- Prepositions *of, at, in, without, between*
- Pronouns *he, they, anybody, it, one*
- Determiners *the, a, that, my, more, much, either, neither*
- Conjunctions *and, that, when, while, although, or*
- Auxiliary verbs *be (is, am, are), have, got, do*
- Particles *no, not, nor, as*

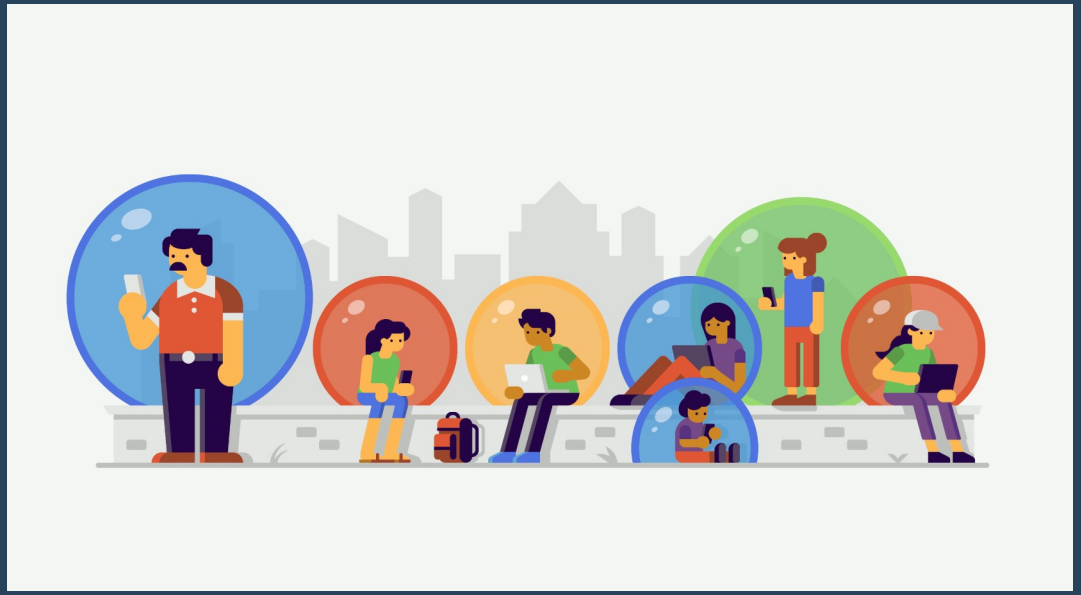
Everyday hallucinations are kind of fine, actually, aren't they?

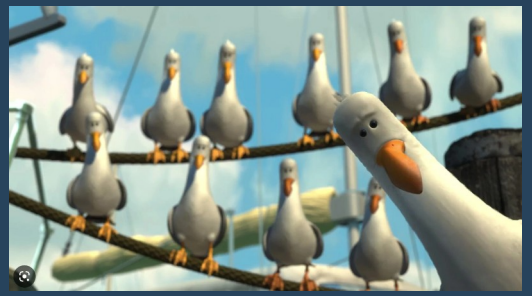
Your brain/mind
is your personal
spaceship and
your personal holodeck.
We like this.
It's better than any
alternative!

(We will need to force our machines to share our
shared human loosely collective consciousness "hallucination" so they can stay in the same "lane of crazy" as we're all already in together, share in this shared reality and this common ground that is accepted as mostly ubiquitous.)

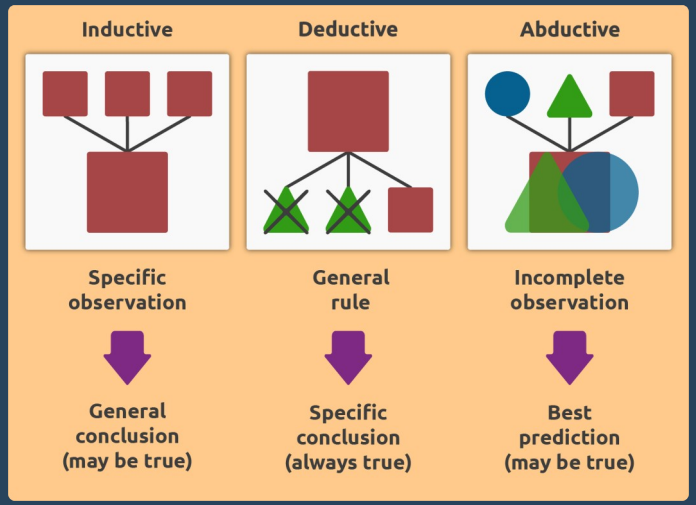
Oops! All HALLUCINATIONS
(everybody, all day,
every day, in some sense)

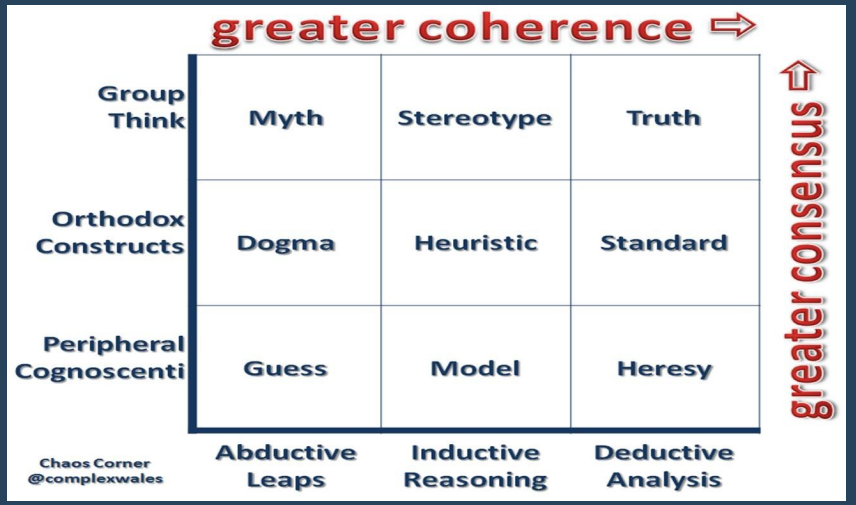
Shared hallucinations: filter bubbles and echo chambers






MANY OTHERS SHARE MY SAME MISCONCEPTIONS
SO I GOT THAT GOIN' FOR ME WHICH IS NICE


Inductive
Deductive
Abductive
Specific observation
General rule
Incomplete observation
General conclusion (may be true)
Specific conclusion (always true)
Best prediction (may be true)


greater coherence ⇨
greater consensus ⇧
Group Think
Orthodox Constructs
Peripheral Cognoscenti
Myth
Stereotype
Truth
Dogma
Heuristic
Standard
Guess
Model
Heresy
Chaos Corner @complexwales
Abductive Leaps
Inductive Reasoning
Deductive Analysis

Pedro Domingos
13K Tweets

Tweets | Tweets & replies | Media | Likes

61 33 217 321K

Promoted

Pedro Domingos @pmddomingos · 15h
AI systems used to do only deduction. Now they do induction even when they should do deduction (a.k.a. hallucinate).

2 21 2,849

Pedro Domingos @pmddomingos · 15h
If you've found the truth, you should welcome scrutiny, not fear it.

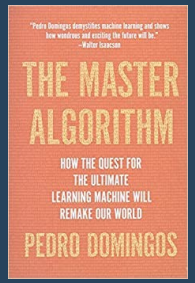
4 9 40 3,141

Pedro Domingos @pmddomingos · 15h
There's only one industry in the world where the product keeps getting worse and the price keeps getting higher: higher education.

27 46 321 13.8K

Pedro Domingos @pmddomingos · 1d
I've been an AI expert for 30 years. These days I meet a lot of people who've been AI experts for 3 months.

212 205 2,086 222K



IBM Watson, Jeopardy winning, etc.

- IBM likes to play their cards close to their vest as to how Watson works
- core developer in TV interview (paraphrasing):
  - “We can’t teach machines what words MEAN, per se”
  - Continuing, with glint in eye…
    - “…but we can teach machines how words are related to other words”
  - Implication: that’s maybe about all you need to “act as if” you know what words mean. Probably not 100% on base, but close to true.
  - Emergent capabilities come from extreme interconnectedness:
    - See also: ChatGPT, and “Attention is all you need”, human associative memory, language learners, early readers

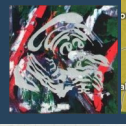
HAYSTACK
The Search Relevance Conference

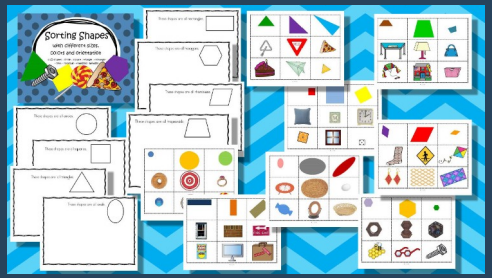
Beyond nearest neighbors…

Sometimes we want little bit of this and a little bit of that

Things do not always fit nicely.

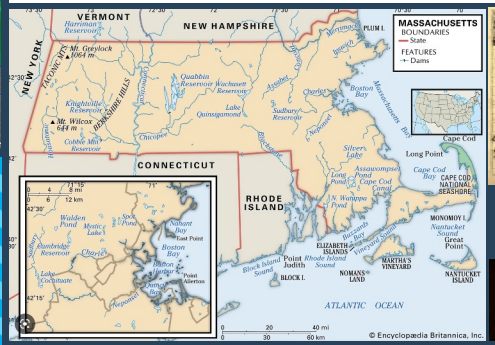
Mix it up!

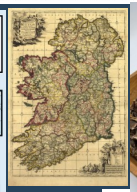


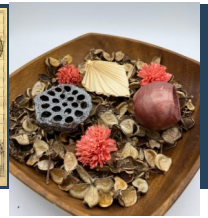


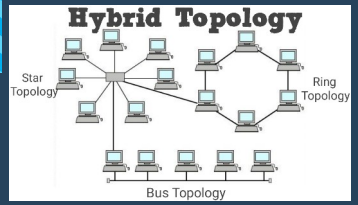
The spice must flow!

BUSINESS: bundling and unbundling







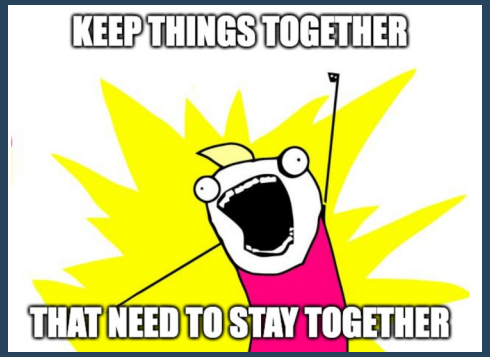


Whatever rules you make, people will want to break them. Nature will break your rules, if only just to remind you who's really "Boss".

Draw multiple lassos around disparate component ingredients and call it your new awesome sauce!

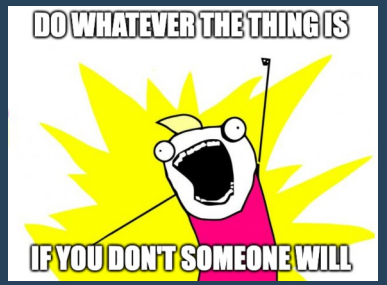
OpenSource Connections

#haystackconf www.haystackconf.com

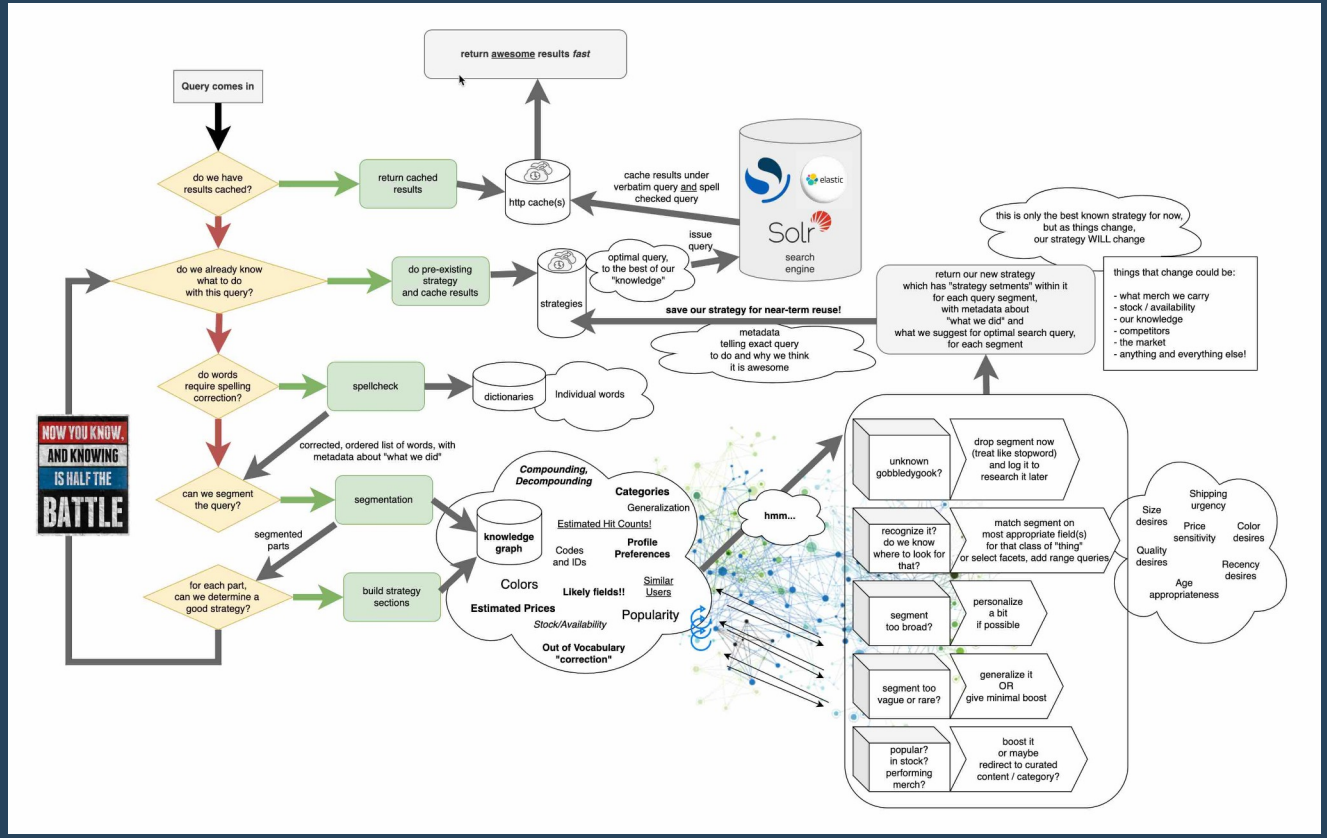

KEEP THINGS TOGETHER
THAT NEED TO STAY TOGETHER


PUT THINGS TOGETHER
THAT DON'T GO TOGETHER


INCLUDE AND/OR EXCLUDE
GET WHATEVER YOU WANT


DO WHATEVER THE THING IS
IF YOU DON'T SOMEONE WILL


Applause


return awesome results fast
Query comes in
do we have results cached?
return cached results
http cache(s)
cache results under verbatim query and spell checked query
elastic
Solr
search engine
issue query
do we already know what to do with this query?
do pre-existing strategy and cache results
strategies
optimal query, to the best of our "knowledge"
save our strategy for near-term reuse!
metadata telling exact query to do and why we think it is awesome
return our new strategy which has "strategy setments" within it for each query segment, with metadata about "what we did" and what we suggest for optimal search query, for each segment
this is only the best known strategy for now, but as things change, our strategy WILL change
things that change could be:
- what merch we carry
- stock / availability
- our knowledge
- competitors
- the market
- anything and everything else!
do words require spelling correction?
spellcheck
dictionaries
Individual words
NOW YOU KNOW, AND KNOWING IS HALF THE BATTLE
corrected, ordered list of words, with metadata about "what we did"
can we segment the query?
segmentation
segmented parts
for each part, can we determine a good strategy?
build strategy sections
Compounding, Decompounding
knowledge graph
Categories
Generalization
Estimated Hit Counts!
Codes and IDs
Profile Preferences
Colors
Likely fields!!
Similar Users
Estimated Prices
Stock/Availability
Popularity
Out of Vocabulary "correction"
hmm...
unknown gobbledygook?
drop segment now (treat like stopword) and log it to research it later
recognize it? do we know where to look for that?
match segment on most appropriate field(s) for that class of "thing" or select facets, add range queries
Size desires
Shipping urgency
Price sensitivity
Color desires
Quality desires
Recency desires
Age appropriateness
segment too broad?
personalize a bit if possible
segment too vague or rare?
generalize it OR give minimal boost
popular? in stock? performing merch?
boost it or maybe redirect to curated content / category?

Want to collaborate at the hackathon tomorrow if you're going to be around!

Let's build cool stuff together!

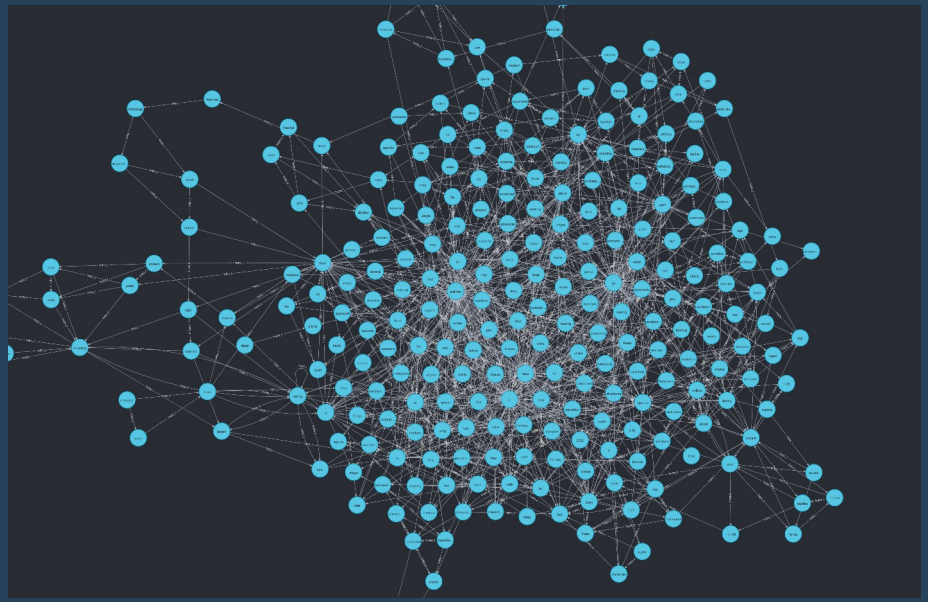
Thank you!

- Time for Q & A / feedback / refreshing anecdotes and/or thinly veiled ad hominem attacks (a joke! please wait until later!)

- Contact Info:

  - Chris Morley of Wellesley, Mass.
  - cmorley@opensourceconnections.com
  - @depahelix on Twitter
  - I'm on Relevance Slack
  - Contact OpenSource Connections!

Example graph I quickly loaded…



Casually from just a few pages off Wikipedia…

Only a few hundred lines of Python using these imports:

nltk, uuid, re, bs4, requests, dictionary, spaCy, pandas py2neo

You can do it too!

Have fun!