Topic Modeling for Conceptual Cartography

FU Berlin Colloquium, 7/12/2022

Charles H. Pence

@pence@scholar.social

UCLouvain

Institut supérieur de philosophie (ISP)

Outline

1. Why Topic Modeling?
2. Basic Topic Modeling
3. Dynamic Topic Modeling
4. Correlating Topics and Features
5. Some Morals

The take-home: Topic modeling *can* be useful for mapping a concept, but we need to be attentive to its failure modes!

Why Topic Modeling?

Topic Models

An **unsupervised** method to reduce a corpus of documents to a smaller collection of **topics** that are **human-interpretable.**

The Usual

Normally taken to be a way in which you can learn **what your corpus is about.** What subjects are discussed, where, and by whom?

The Usual

Normally taken to be a way in which you can learn **what your corpus is about.** What subjects are discussed, where, and by whom?

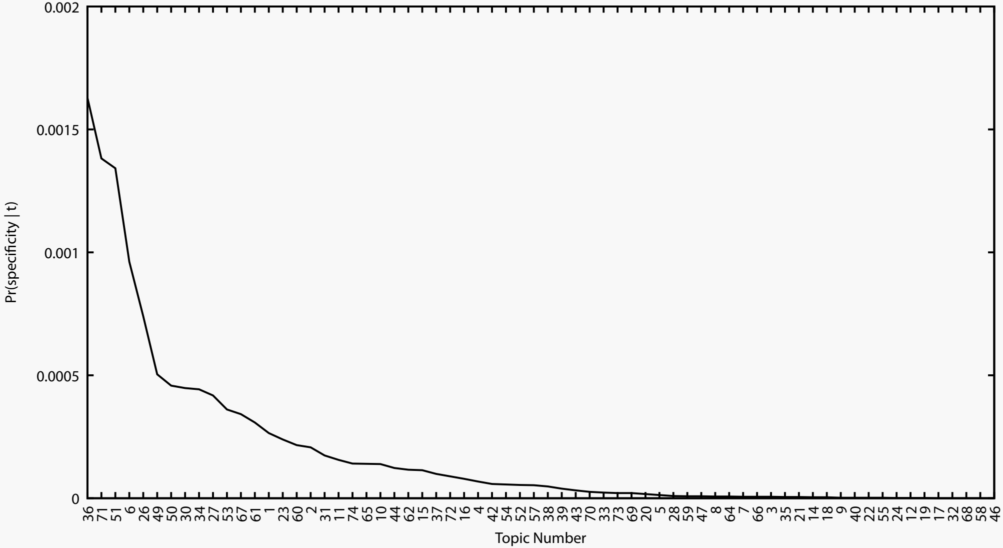
My goal today: Can we use the same idea to **understand the content, nature, and change of concepts** across a corpus?

Basic Topic Modeling

Case Study

How should we understand the concept of **specificity** in the life sciences?

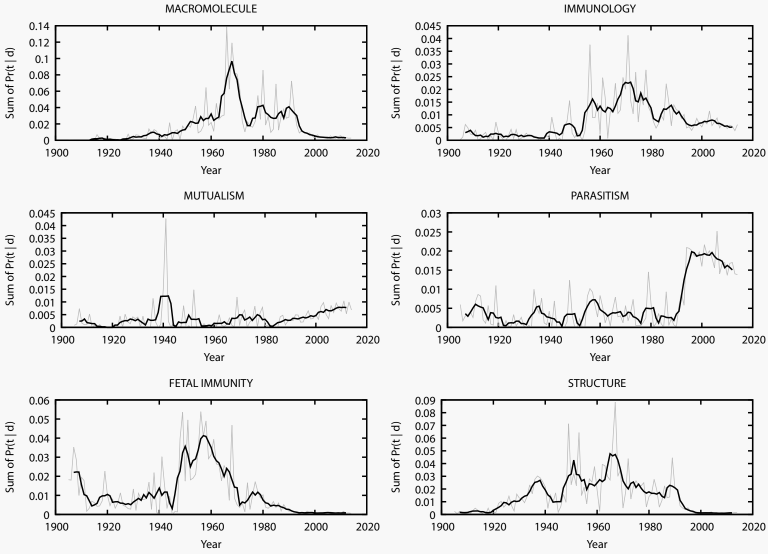
Topics with ‘Specificity’



Topics with ‘Specificity’

Take the top six of those topics and look at their evolution over time, as a proxy for different **senses** of the term in the literature.

Topics with ‘Specificity’



Challenges

Of course, topics involve **lots** of probable words! So we're not looking at **definitions** of a concept so much as **contexts of usage of a term.** Question: What can those teach us?

Challenges

Of course, topics involve **lots** of probable words! So we're not looking at **definitions** of a concept so much as **contexts of usage of a term.** Question: What can those teach us?

Also: What to do with concepts that go by multiple names?

Dynamic Topic Modeling

Dynamic Topic Models

In a normal topic model, the probability for a word in a topic is **fixed across the corpus.**

Dynamic topic models: divide the corpus into chunks, here corresponding to time-periods, and **allow those probabilities to vary** (Blei and Lafferty 2006).

Dynamic Topic Models

Intuitively: a way to say that some topic is **the same topic** over time, while particular words become more or less important for that topic.

Or, following my project here: to track shifting conceptual commitments within a field?

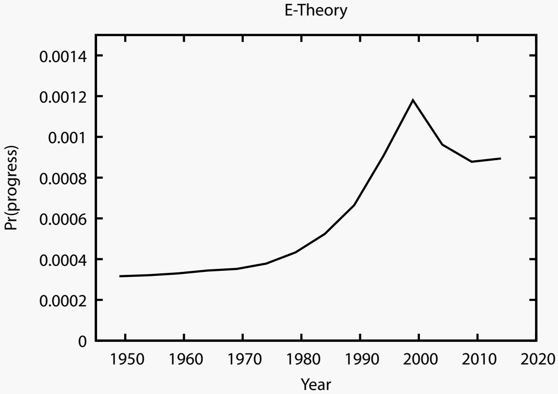
A Case Study

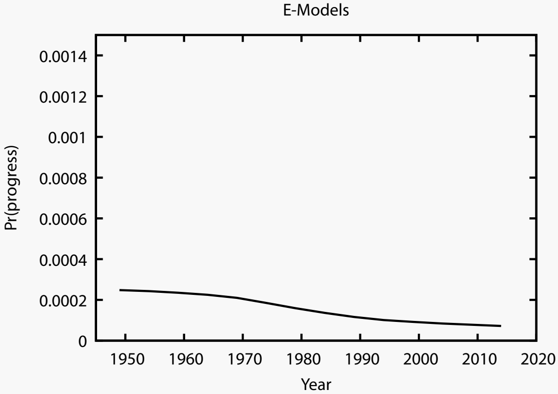
The concept of **progress** in evolution — explored through the journal *Evolution*

Progress in Evolution

Two topics that have non-zero probabilities for ‘progress’:

- E-Theory (13): Prior to 1970, picks out theoretical papers in evolutionary biology; then especially book reviews (as the “most theoretical” content in the journal); then public-facing
- E-Models (17): Formal modeling results in evolutionary theory


E-Theory
Pr(progress)
0.0014
0.0012
0.001
0.0008
0.0006
0.0004
0.0002
0
1950
1960
1970
1980
1990
2000
2010
2020
Year


E-Models
Pr(progress)
0.0014
0.0012
0.001
0.0008
0.0006
0.0004
0.0002
0
1950
1960
1970
1980
1990
2000
2010
2020
Year

Changes in Words

E-Theory, 1949 vs. 1979:

Increasing Words	Decreasing Words
book: +0.003514	time: -0.001870
theory: +0.002712	primitive: -0.001693
chapter: +0.002214	know: -0.001582
evolutionary: +0.001942	genera: -0.001557
biology: +0.001718	rodent: -0.001523
	(...)
	man: -0.001212
	modern: -0.000710

Challenges

- Disentangling changes in **topic assignation** from changes in **topic content**
- Interpreting the **disappearance** of something from the corpus

Correlating Topics and Features

Taxonomy Corpus

A corpus of around 40,000 articles in **biological taxonomy.**

Idea: What if we correlate the presence of particular **features** in the documents (like reference to different species, or to different concepts of what a “species” is) to topics?

Topic-Feature Correlation

Topic 16: popular in mammals

- 0.027*"colombia"
- 0.016*"specie"
- 0.013*"type"
- 0.013*"peru"
- 0.010*"locality"
- 0.010*"venezuela"
- 0.010*"ecuador"
- 0.009*"panama"
- 0.008*"distribution"
- 0.007*"brazil"
- 0.007*"key"
- 0.006*"rica"
- 0.006*"del"
- 0.006*"costa"
- 0.006*"genus"
- 0.006*"male"
- 0.006*"america"
- 0.006*"san"
- 0.006*"neotropical"
- 0.005*"cat"

Topic-Feature Correlation

Topic 16: popular in mammals

- 0.027*"colombia"
- 0.016*"specie"
- 0.013*"type"
- 0.013*"peru"
- 0.010*"locality"
- 0.010*"venezuela"
- 0.010*"ecuador"
- 0.009*"panama"
- 0.008*"distribution"
- 0.007*"brazil"
- 0.007*"key"
- 0.006*"rica"
- 0.006*"del"
- 0.006*"costa"
- 0.006*"genus"
- 0.006*"male"
- 0.006*"america"
- 0.006*"san"
- 0.006*"neotropical"
- 0.005*"cat"

Okay: Central and South American collection sites

Interesting Correlations

Topic 9: traditional specimen collection terms

- 0.029*"specie"
- 0.012*"forest"
- 0.012*"habitat"
- 0.010*"area"
- 0.008*"find"
- 0.007*"collect"
- 0.007*"site"
- 0.007*"study"
- 0.007*"record"
- 0.006*"population"
- 0.006*"range"
- 0.006*"high"
- 0.005*"specimen"
- 0.005*"occur"
- 0.005*"know"
- 0.004*"individual"
- 0.004*"region"
- 0.004*"number"
- 0.004*"sample"
- 0.004*"distribution"

Popular in every taxon **except** non-insect arthropods, fish, and fungi.

Interesting Correlations

Topic 64: molecular phylogenetics

- 0.021*"specie"
- 0.017*"sequence"
- 0.016*"analysis"
- 0.011*"molecular"
- 0.010*"dna"
- 0.008*"phylogenetic"
- 0.007*"tree"
- 0.007*"clade"
- 0.007*"gene"
- 0.007*"specimen"
- 0.007*"study"
- 0.007*"morphological"
- 0.006*"support"
- 0.006*"group"
- 0.006*"genetic"
- 0.006*"coi"
- 0.006*"datum"
- 0.006*"base"
- 0.005*"table"
- 0.005*"population"

Among the **top-20 most significant probabilities** in reptiles and amphibia, birds, fish, fungi, and mammals; top-5% in every other group

Troublesome Correlations

Topic 31:

- 0.016*"male"
- 0.016*"genitalia"
- 0.013*"specie"
- 0.009*"female"
- 0.009*"fig"
- 0.008*"brown"
- 0.008*"lepidoptera"
- 0.007*"scale"
- 0.007*"long"
- 0.006*"slide"
- 0.006*"white"
- 0.006*"line"
- 0.006*"new"
- 0.006*"bursae"
- 0.006*"short"
- 0.005*"dark"
- 0.005*"coll"
- 0.005*"forewing"
- 0.005*"holotype"
- 0.005*"leg"

Cautious hypothesis: Lepidopteran anatomy, especially reproductive

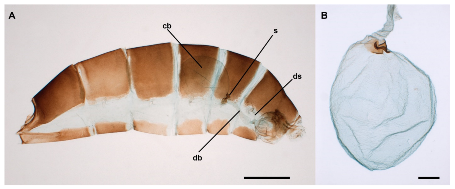
Troublesome Correlations

But wait.

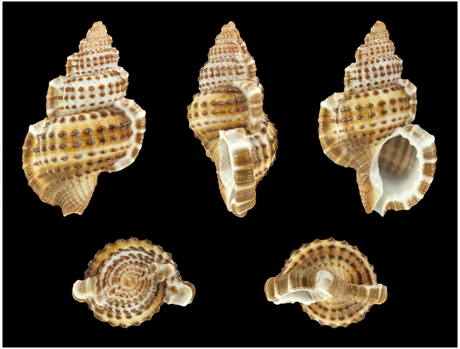
Our lepidopteran reproductive anatomy topic is unusually significant in one group... **in papers that mention molluscs.**

...what?

One Hypothesis?



bursa copulatrix, *Leptophobia aripa*



genus *Bursa*, *Bursa granularis*

Boring Correlations

- Topic 22 (fish anatomy): prevalent in fish
- Topic 32 (reptile anatomy): prevalent in reptiles, amphibians, fish
- Topic 83 (beetle anatomy): prevalent in insects

Even More Boring Anti-Correlations

- Topic 2 (insects/worms): anti-correlated with fish
- Topic 11 (jewel beetles): anti-correlated with mammals

Challenges

- Is there some way to **sort** the boring stuff from the non-boring stuff? (Lots of classic significance tests don't seem to do it.)
- Can we recover useful **anti-correlations** or are they doomed to be boring?

Some Morals?

Some Morals?

Getting from **text** to **concepts** will of course never be easy – I've ignored a variety of issues in linguistics here as well.

What are the uses of the kind of **cartography** that we can do in these contexts? How can we best put it in dialogue with traditional close reading?

Questions?

charles@charlespence.net
https://pencelab.be
@pence@scholar.social