arXiv recommendations

providing similar articles from journal abstract analysis

overview

Problem:
Scientists have limited time but need to keep on top of current research.

Project:
I've created a recommendation system that suggests similar abstracts to help streamline research.

data: source

arXiv offerings and issues:

- API offers XML downloads of article metadata including title, abstract, authors, categories
- API is flaky and returns a lot of empty calls
- XML is more nested than xmltodict can handle

unnest the data:

- user-created Python module arXivpy downloads and unnests most columns
- not a perfect solution; doesn't notify user when arXiv returns empty API call

data: the dataset

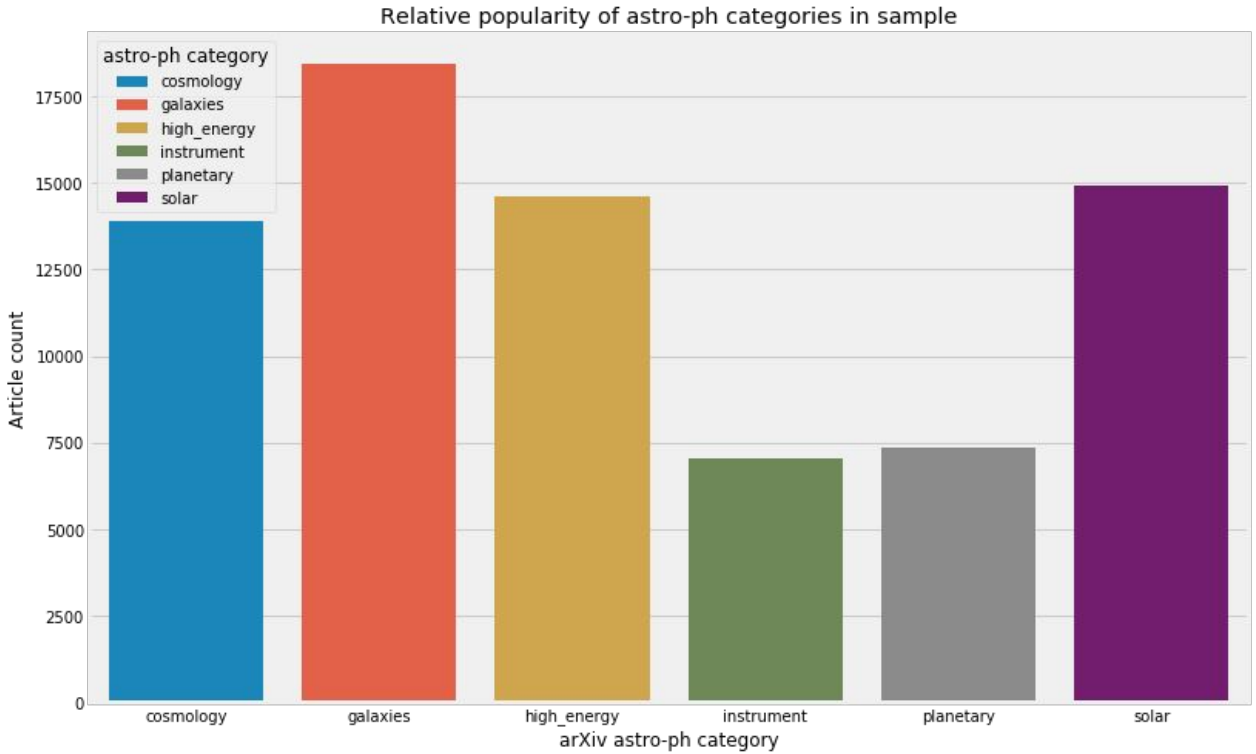
filtered out:

- abstract exact duplicates
- abstracts that “have been withdrawn” by the author
- primary category is the deprecated “astro-ph”

final:

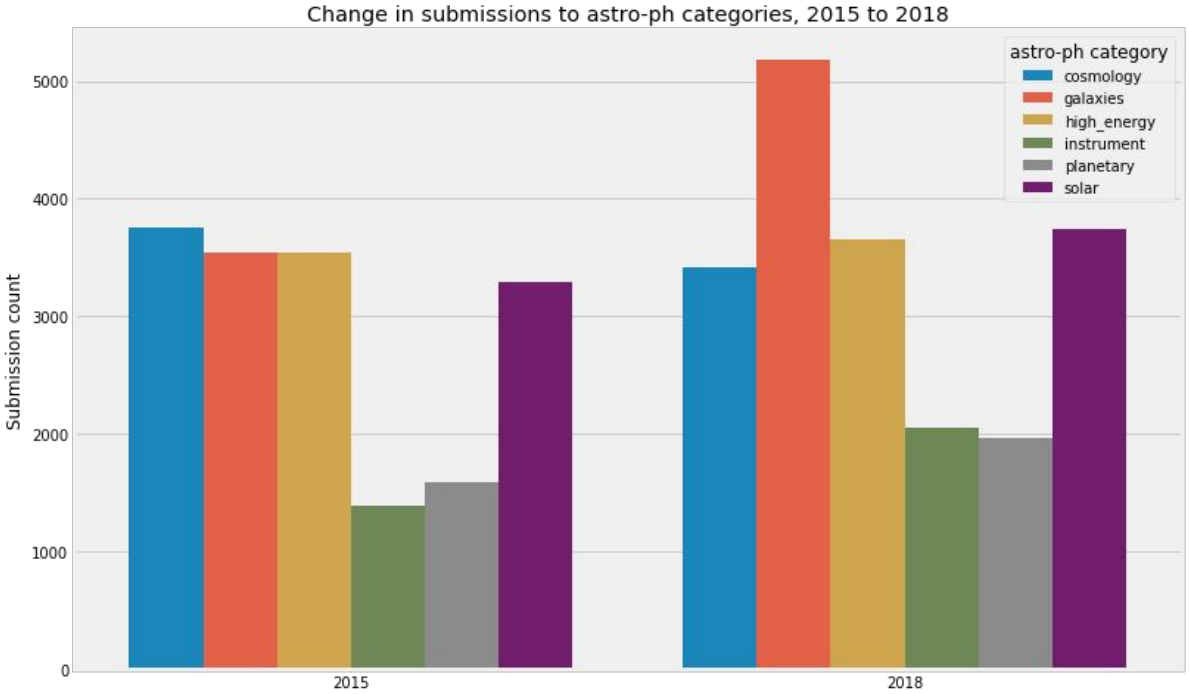
- 59,972 abstracts
- covers ~4 years
- primary category might be outside astrophysics
- one article submitted to all six astro-ph categories

astro-ph categories



- most popular: astro-ph.GA (galaxies), 15.9k articles
- least popular: astro-ph.IM (instruments), 7k
- average: 12.7k

did category popularity change after LIGO?



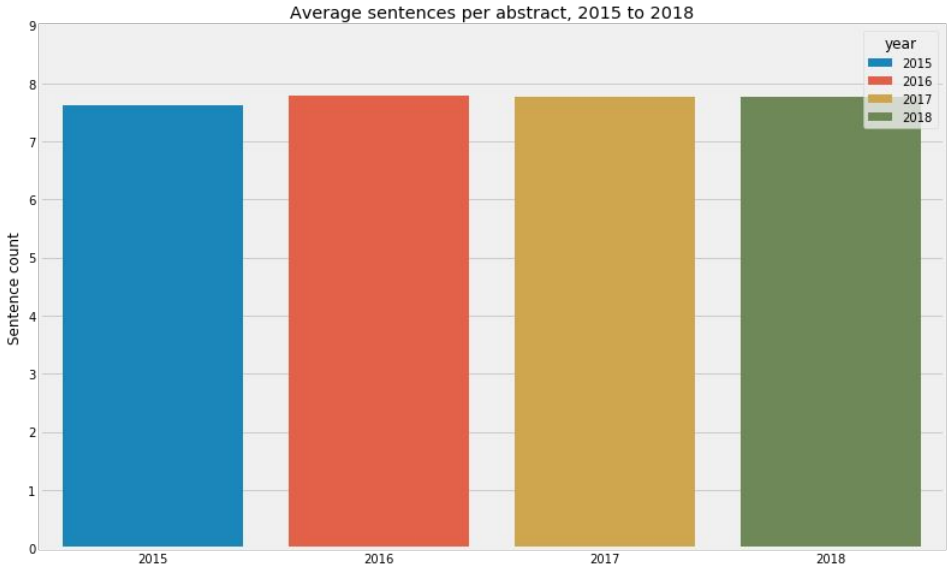
- overall submissions +17%, 2015 to 2018
- **galaxies +46%**
- **instruments +46%**
- planetary +23%
- solar +13%
- high energy +3%
- cosmology -9%

sentences

2015 is weird

- statistical testing shows a difference between 2015 and 2016-2018
- 2015: 7.6 sentences per abstract
- 2018: 7.8 sentences per abstract

¯_(ツ)_/¯



cleaning

premodeling text cleaning

issue: LaTeX

- special mathematical formatting popular with scientists
- difficult to remove
- solution: regular expressions
    - ‘\$\S*\$’ removes $math$
    - some had to be hard-coded for removal

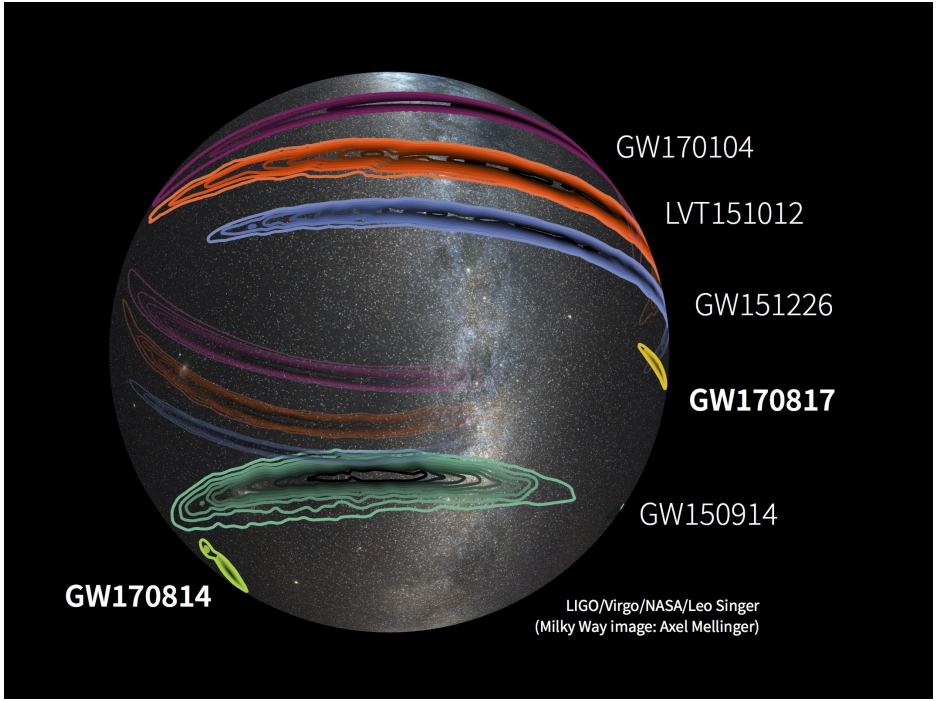
issue: LaTeX leftovers

- single letters wandered freely
- forgot about “x-ray”
    - removing remaining punctuation removed important word
    - solution: converted “x-” to “x_” and removed only non-underscore punctuation

premodeling text cleaning

issue: numbers

- astrophysicists like to name stuff
- ... with numbers
- can I remove digits safely?
- testing: only one named event came close to threshold
- decision: remove digits, save processor



modeling

model: tfidf

“term frequency-inverse document frequency”

1) how frequently does this word appear in *this* abstract?
2) how frequently does it appear in *all* of my abstracts?

word is more important when it appears in fewer abstracts

tfidf: parameters

vocabulary

- add word to vocabulary if in at least 400 abstracts
- as number goes lower, typos and mistakes creep in
- balancing act between uncommon words and common mistakes

“ngrams”

- groups of words that appear together at least 400 times
- created phrases up to 3 words
- bigrams: “massive star,” “binary star,” “milky way,” “giant planet”
- trigram: “hubble space telescope”

recommendations

essentials

proof-of-concept:

trained on 98% of data -- 1,200 abstracts not modeled are “user input”

system:

tfidf calculates similarity between abstracts based on important words

input:

random abstract that model never saw (unknown words in abstract?)

output:

list of abstracts with highest similarity

output: first checks

	abstract	title	terms	document_similarity
31861	Nearby star-forming galaxies offer a unique en...	Different generations of HMXBs: clues about th...	astro-ph.HE	0.503205
2748	We have identified 55 candidate high-mass X-ra...	Formation Timescales for High-Mass X-ray Binar...	astro-ph.HE|astro-ph.GA	0.479234
1955	[abridged] How does a star cluster of more tha...	NGC 346: Looking in the Cradle of a Massive St...	astro-ph.GA	0.455667
23459	We present 15 high mass X-ray binary (HMXB) ca...	Young Accreting Compact Objects in M31: The Co...	astro-ph.HE	0.446389
23715	The 30 Doradus star-forming region in the Larg...	An excess of massive stars in the local 30 Dor...	astro-ph.SR|astro-ph.GA	0.440245
28188	The objective of this work is to study how act...	Quenching by gas compression and consumption: ...	astro-ph.GA|astro-ph.CO	0.431887

check 1:

do items in “terms” column match?

example has mix of categories;
no category represented on every row

check 2:

do other words resolve confusion?

row 1: “galaxies”
row 4: “M31” (a galaxy)
rows 2, 3, 5, 6: astro-ph.GA

output: more checks

check 3:

do these words fit together? how many important words are in both abstracts?

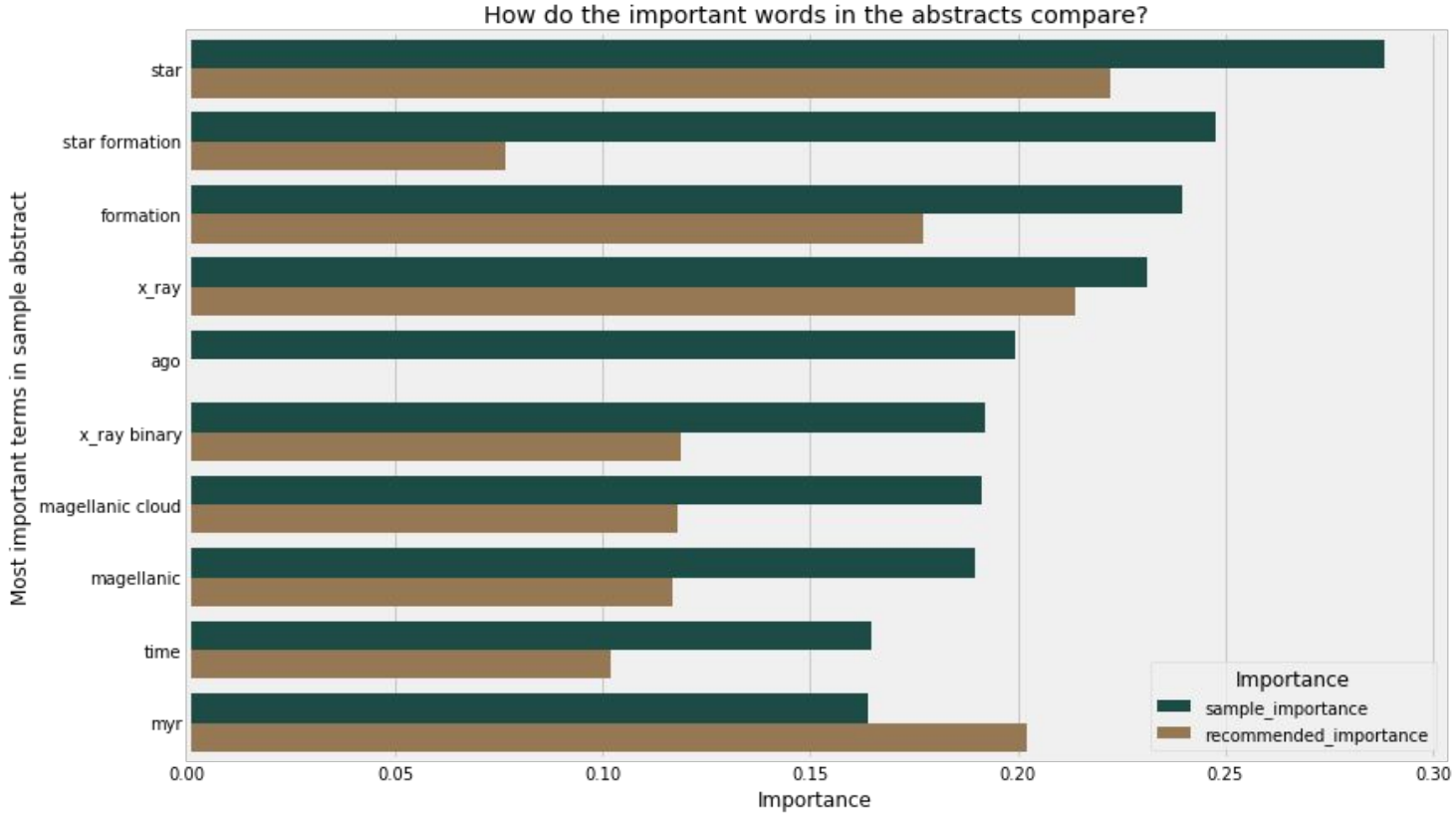
> “ago,” “time,” and “myr” (million years) related

note:

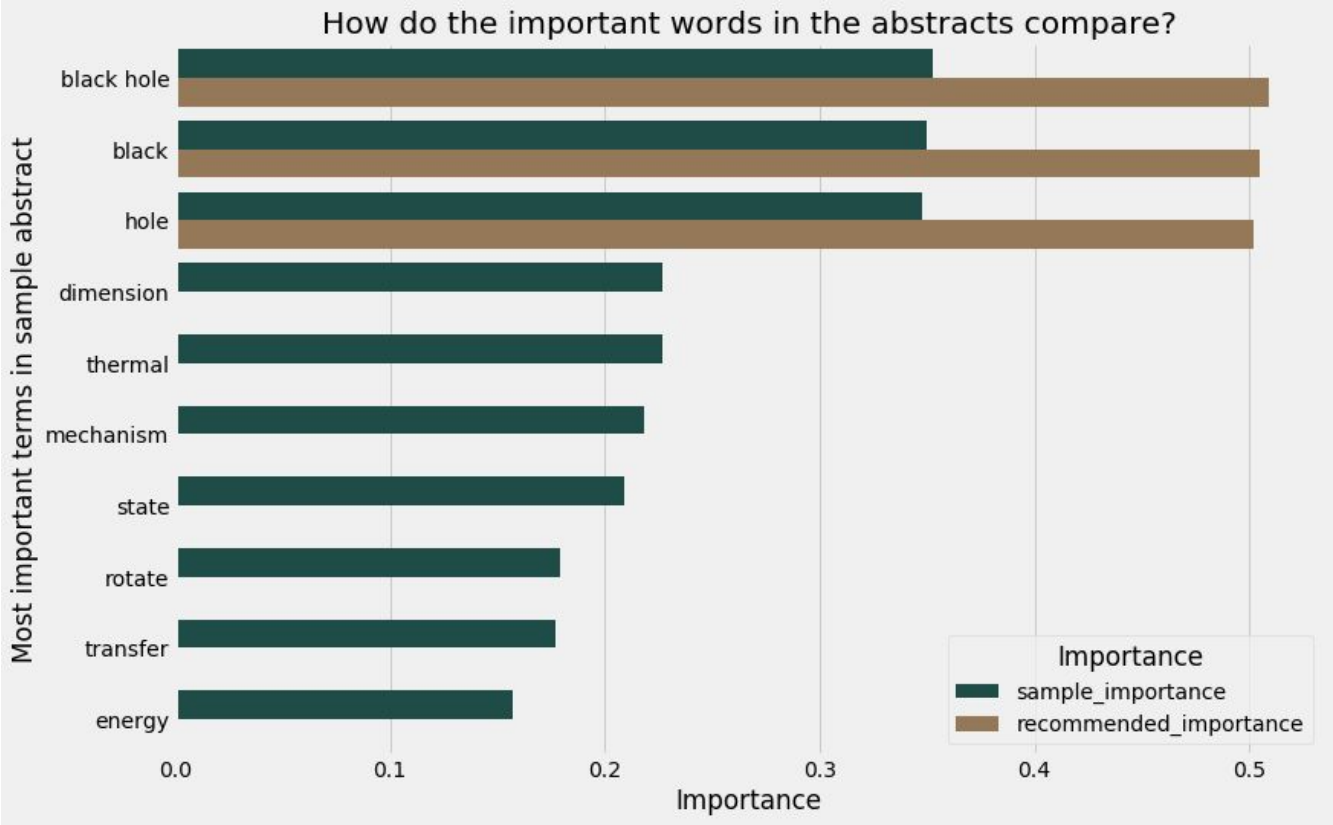
most tables will have fewer matching words

	feature	sample_importance	recommended_importance
1986	star	0.288187	0.222307
1991	star formation	0.247674	0.076422
793	formation	0.239517	0.177373
2272	x_ray	0.230971	0.213804
58	ago	0.199327	0.000000
2273	x_ray binary	0.192241	0.118635
1204	magellanic cloud	0.191314	0.118063
1203	magellanic	0.189575	0.116991
2128	time	0.164884	0.101753
1355	myr	0.163825	0.202199

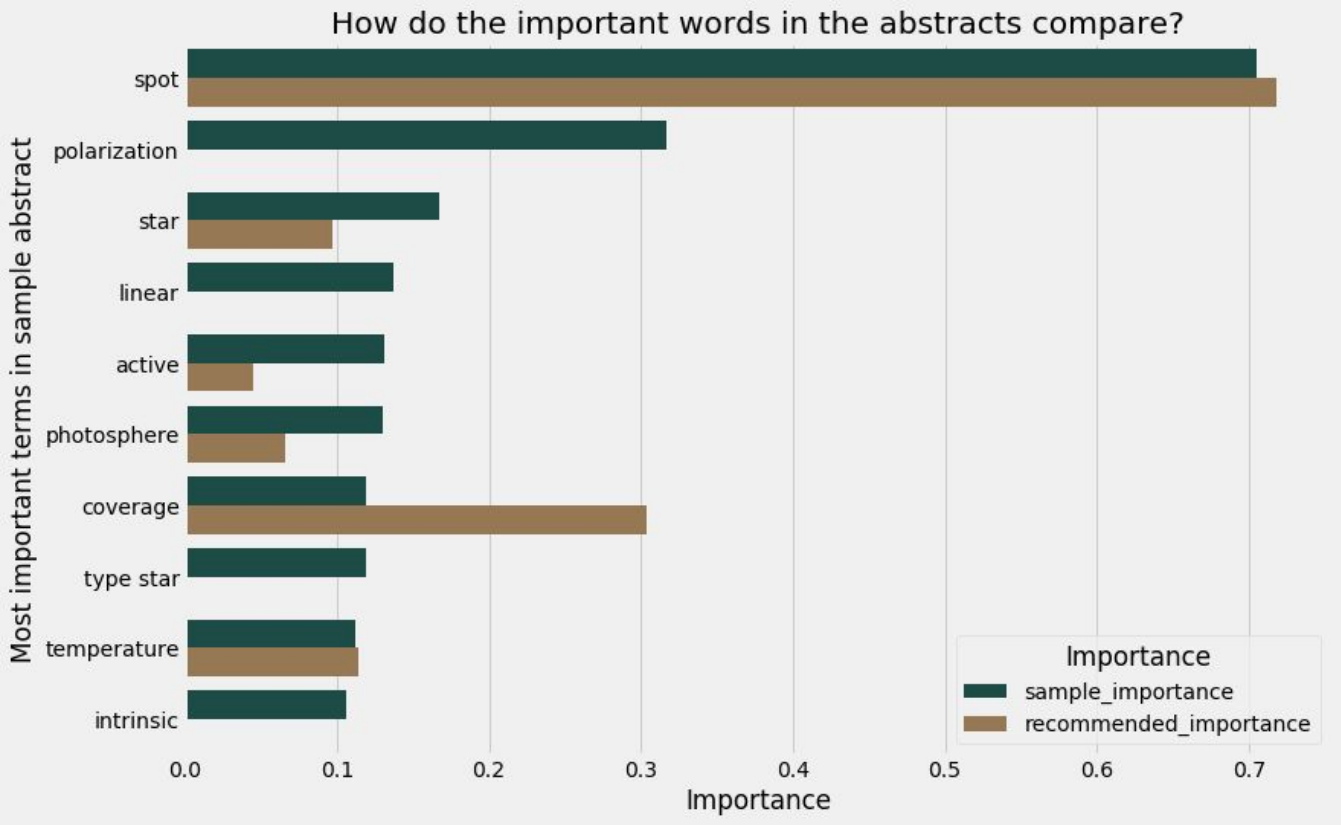
output: features



output: features example 1



output: features example 2



future directions

1. vocabulary refinements (probably regex rewrite)
2. topic modeling/visualizations (how have interests of community changed over time?)
3. allow alphanumeric tokens and word-based search (neural network and/or RAM upgrade)