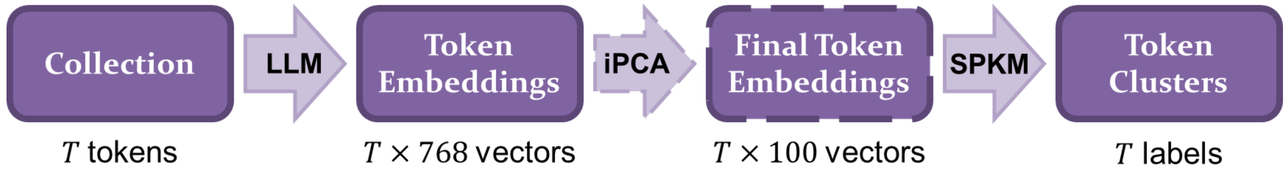
What do contextualized representations represent?

Laure Thompson
laurejt@umass.edu

David Mimno
mimno@cornell.edu

Don't cluster the vocabulary, cluster the tokens!



Clusters capture theme & polysemy like LDA

Term	Model	Top Words
land	LDA	sea coast Beach Point coastal ***land*** Long Bay m sand beach tide Norfolk shore Ocean
		land acres County ha facilities State location property acre cost lot site parking
	BERT	arrived arrival landing landed arriving arrive returning settled departed ***land*** leaving
		land property rights estate acres lands territory estates properties farm farmland
	GPT-2	arrived landed arriving landfall arrive arrives arrival landing ***land*** departed ashore
		land sea ice forest rock mountain ground sand surface beach ocean soil hill lake snow
metal	LDA	metals ***metal*** potassium sodium + lithium compounds electron ions hydrogen
		metal folk bands music genre band debut Metal heavy musicians lyrics instruments
	BERT	metals elements ***metal*** electron element atomic periodic electrons chemical atoms
		rock dance pop ***metal*** Rock folk jazz punk comedy Dance heavy funk alternative soul
	GPT-2	rock pop hop dance ***metal*** folk hip punk jazz B soul funk alternative rap heavy disco
		plutonium hydrogen carbon sodium potassium ***metal*** lithium uranium oxygen

...but are more syntactically aware than LDA

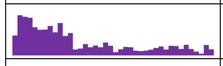
Model	$P_{k\%}$	H	Top Words (noun verb adj adv other)
LDA	5%	0.69	Valley Death valley Creek California mining ° Range Nevada Desert
	25%	0.97	army forces soldiers campaign troops captured defeated Battle victory commander
	50%	1.11	society News Week Good Spirit Fruit says Doug host free
	75%	1.28	Washington Delaware ceremony Grand Capitol building 156 Number laying Master
	95%	1.53	critics reviews review positive mixed list Entertainment Times style something
BERT	5%	0.00	1997 1996 1995 1937 1895 1935 96 1896 1795 97
	25%	0.61	Jewish Israel Jews Ottoman Arab Muslim Israeli Islamic Jerusalem Islam
	50%	0.86	captured defeated attacked capture attack siege destroyed surrender defeat occupied
	75%	1.09	hop dance hip B R Dance Hip Z Hop rapper
	95%	1.48	separate combined co joint shared divided common combination distinct respective
GPT-2	5%	0.00	2004 2003 2015 2000 2014 1998 2001 2013 2002 1997
	25%	0.42	Atlantic Pacific Gulf Mediterranean Caribbean Columbia Indian Baltic Bay Florida
	50%	0.73	knew finds discovers learned reveals discovered know heard discover learns
	75%	1.02	Olympic League FA Summer Premier Division UEFA European Winter Tour
	95%	1.42	positive mixed critical negative garnered favorable mostly attracted commercial

Transfer Learning Supports Corpus Exploration

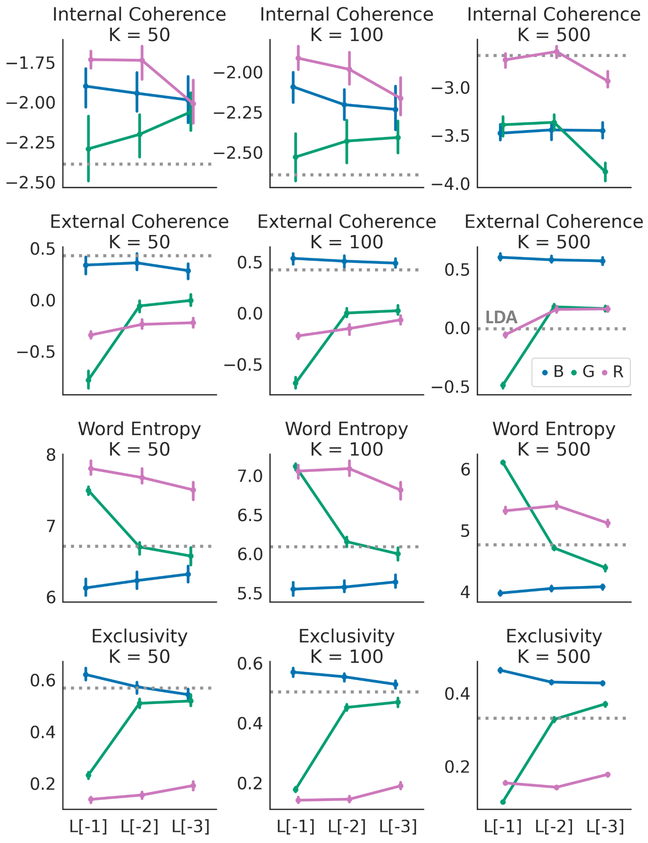
Product Reviews Over Categories

Category	Top words
books	book books author novel novels work Book fiction by authors
	read reading copy Read reads Reading readable reader reread
	problem children problems course power lives mystery questions
electronics	use up than off used back over using there about work need down
	screen quality sound device power battery unit system software
	setup remote battery mode card set range input signal support
movies	movie movies films flick theater Movie flicks game cinema film
	into up over through between off down than about around during
	film movie picture screen documentary films Film cinema feature
cds	album albums record release Album LP releases records effort up
	songs tracks hits tunes singles material stuff Songs ballads cuts
	lyrics guitar vocals voice bass singing solo vocal sound work music

SCOTUS Opinions Over Time

1980–2019	Top words
	union employment labor bargaining Labor workers job strike unions working hiring jobs Employment worker
	gas coal oil natural mining mineral fuel mine fishing hunting timber Coal submerged uranium surface
	compensation wages pension wage salary welfare compensate salaries retirement bonus Pay compensated
	discrimination prejudice unfair bias harassment segregation retaliation boycott persecution unfairness
	medical health care hospital physician patient Medical physicians clinic hospitals doctor surgical doctors
	market competition competitive markets compete demand marketplace trading competitor trade fixing
	election vote voting electoral ballot voter elected votes elect Election districting referendum voted polling
	violence firearm gun violent weapon firearms armed weapons arms lethal guns deadly Violence terrorism
	patent copyright Copyright Patent patents trademark invention patented copyrighted patentee patentable

Topical Quality Varies by Model & Layer



To learn more, check out our preprint on arXiv