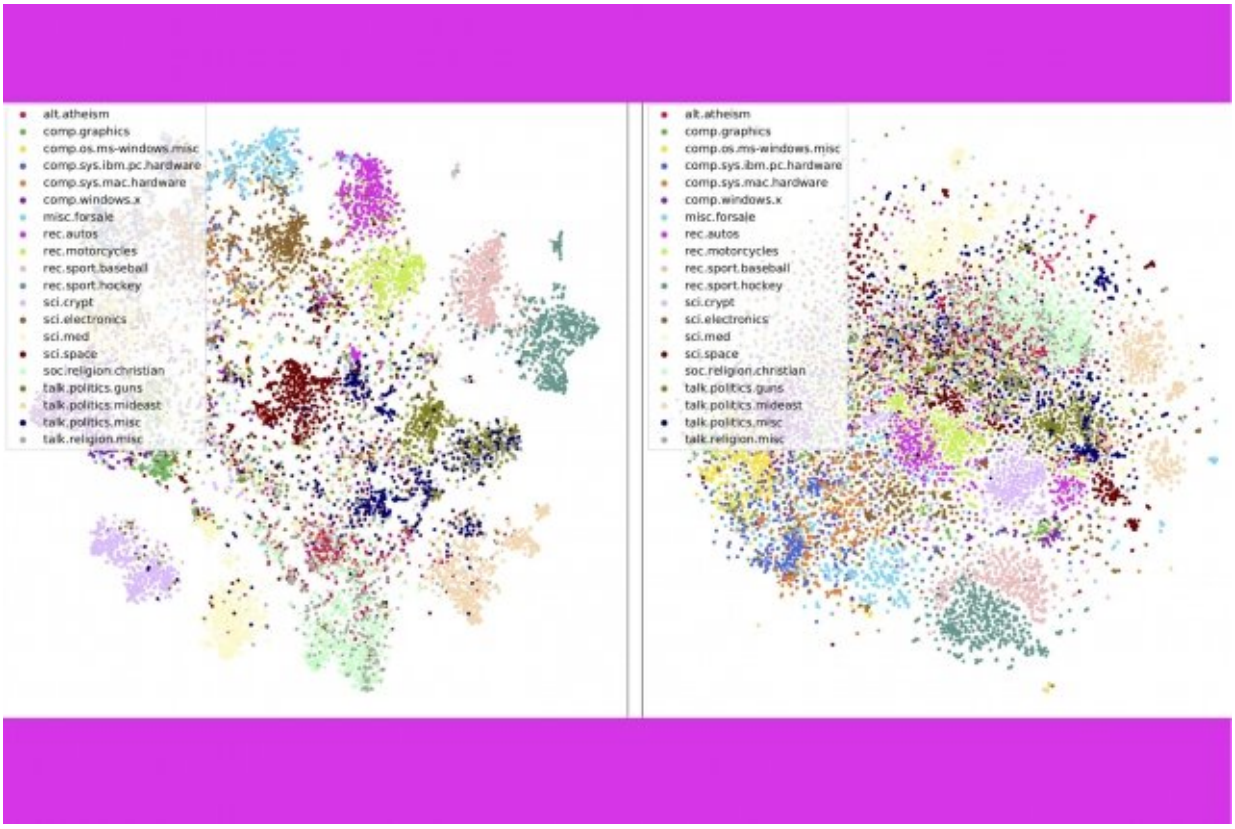
As natural language processing techniques improve, suggestions are getting speedier and more relevant

December 23 2019, by Kim Martineau



In a new study, researchers at MIT and IBM combine three popular text-analysis tools — topic modeling, word embeddings, and optimal transport — to compare thousands of documents per second. Here, they show that their method (left) clusters newsgroup posts by category more tightly than a competing method. Credit: Massachusetts Institute of Technology

With billions of books, news stories, and documents online, there's never been a better time to be reading—if you have time to sift through all the options. "There's a ton of text on the internet," says Justin Solomon, an assistant professor at MIT. "Anything to help cut through all that material is extremely useful."

With the MIT-IBM Watson AI Lab and his Geometric Data Processing Group at MIT, Solomon recently presented a new technique for cutting through massive amounts of text at the Conference on Neural Information Processing Systems (NeurIPS). Their method combines three popular text-analysis tools—topic modeling, word embeddings, and optimal transport—to deliver better, faster results than competing methods on a popular benchmark for classifying documents.

If an algorithm knows what you liked in the past, it can scan the millions of possibilities for something similar. As natural language processing techniques improve, those "you might also like" suggestions are getting speedier and more relevant.

In the method presented at NeurIPS, an algorithm summarizes a collection of, say, books, into topics based on commonly-used words in the collection. It then divides each book into its five to 15 most important topics, with an estimate of how much each topic contributes to the book overall.

To compare books, the researchers use two other tools: word embeddings, a technique that turns words into lists of numbers to reflect their similarity in popular usage, and optimal transport, a framework for calculating the most efficient way of moving objects—or data points—among multiple destinations.

Word embeddings make it possible to leverage optimal transport twice: first to compare topics within the collection as a whole, and then, within

any pair of books, to measure how closely common themes overlap.

The technique works especially well when scanning large collections of books and lengthy documents. In the study, the researchers offer the example of Frank Stockton's "The Great War Syndicate," a 19th century American novel that anticipated the rise of nuclear weapons. If you're looking for a similar book, a topic model would help to identify the dominant themes shared with other books—in this case, nautical, elemental, and martial.

But a topic model alone wouldn't identify Thomas Huxley's 1863 lecture, "The Past Condition of Organic Nature," as a good match. The writer was a champion of Charles Darwin's theory of evolution, and his lecture, peppered with mentions of fossils and sedimentation, reflected emerging ideas about geology. When the themes in Huxley's lecture are matched with Stockton's novel via optimal transport, some cross-cutting motifs emerge: Huxley's geography, flora/fauna, and knowledge themes map closely to Stockton's nautical, elemental, and martial themes, respectively.

Modeling books by their representative topics, rather than individual words, makes high-level comparisons possible. "If you ask someone to compare two books, they break each one into easy-to-understand concepts, and then compare the concepts," says the study's lead author Mikhail Yurochkin, a researcher at IBM.

The result is faster, more accurate comparisons, the study shows. The researchers compared 1,720 pairs of books in the Gutenberg Project dataset in one second—more than 800 times faster than the next-best method.

The technique also does a better job of accurately sorting documents than rival methods—for example, grouping books in the Gutenberg

dataset by author, product reviews on Amazon by department, and BBC sports stories by sport. In a series of visualizations, the authors show that their method neatly clusters documents by type.

In addition to categorizing documents quickly and more accurately, the method offers a window into the model's decision-making process. Through the list of topics that appear, users can see why the model is recommending a document.

More information: Hierarchical Optimal Transport for Document Representation. arXiv:1906.10827v2 [cs.LG]: arxiv.org/pdf/1906.10827.pdf

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: As natural language processing techniques improve, suggestions are getting speedier and more relevant (2019, December 23) retrieved 10 May 2024 from https://techxplore.com/news/2019-12-natural-language-techniques-speedier-relevant.html