New algorithm can separate unstructured text into topics with high accuracy and reproducibility

January 29 2015, by Emily Ayshford

Luís Amaral

Much of our reams of data sit in large databases of unstructured text. Finding insights among emails, text documents, and websites is extremely difficult unless we can search, characterize, and classify their text data in a meaningful way.

One of the leading big data algorithms for finding related topics within unstructured text (an area called topic modeling) is latent Dirichlet allocation (LDA). But when Northwestern University professor Luis

Amaral set out to test LDA, he found that it was neither as accurate nor reproducible as a leading topic modeling algorithm should be.

Using his network analysis background, Amaral, professor of chemical and biological engineering in Northwestern's McCormick School of Engineering and Applied Science, developed a new topic modeling algorithm that has shown very high accuracy and reproducibility during tests. His results, published with co-author Konrad Kording, associate professor of physical medicine and rehabilitation, physiology, and applied mathematics at Northwestern, were published Jan. 29 in *Physical Review X*.

Topic modeling algorithms take unstructured text and find a set of topics that can be used to describe each document in the set. They are the workhorses of big data science, used as the foundation for recommendation systems, spam filtering, and digital image processing. The LDA topic modeling algorithm was developed in 2003 and has been widely used for academic research and for commercial applications, like search engines.

When Amaral explored how LDA worked, he found that the algorithm produced different results each time for the same set of data, and it often did so inaccurately. Amaral and his group tested LDA by running it on documents they created that were written in English, French, Spanish, and other languages. By doing this, they were able to prevent text overlap among documents.

"In this simple case, the algorithm should be able to perform at 100 percent accuracy and reproducibility," he said. But when LDA was used, it separated these documents into similar groups with only 90 percent accuracy and 80 percent reproducibility. "While these numbers may appear to be good, they are actually very poor, since they are for an exceedingly easy case," Amaral said.

To create a better algorithm, Amaral took a network approach. The result, called TopicMapping, begins by preprocessing data to replace words with their stem (so "star" and "stars" would be considered the same word). It then builds a network of connecting words and identifies a "community" of related words (just as one could look for communities of people in Facebook). The words within a given community define a topic.

The algorithm was able to perfectly separate the documents according to language and was able to reproduce its results. It also had high accuracy and reproducibility when separating 23,000 scientific papers and 1.2 million Wikipedia articles by topic.

These results show the need for more testing of big data algorithms and more research into making them more accurate and reproducible, Amaral said.

"Companies that make products must show that their products work," he said. "They must be certified. There is no such case for algorithms. We have a lot of uninformed consumers of big data algorithms that are using tools that haven't been tested for reproducibility and accuracy."

Provided by Northwestern University

Citation: New algorithm can separate unstructured text into topics with high accuracy and reproducibility (2015, January 29) retrieved 20 September 2024 from https://phys.org/news/2015-01-algorithm-unstructured-text-topics-high.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.