Evolution of Topics and Novelty in Science

GTM 2019 Extended Abstract

Topic models can provide an insight into the semantic structures of texts. These techniques, originated in computer science have emerged as a possible solution to topic extraction from scientific publications. In this paper, we implement Doc2Vec, one of the most recent developments in topic modelling based on Neural Networks, on the production of researchers. We overcome the major limitations described in the literature of the application of topic models to knowledge domains, and explore how our model helps discover novelty and interdisciplinary research.

Introduction

Increases in computational capacity and the availability of (electronic) data have opened many new avenues for estimating document similarity and carrying out clustering. While the range of options and ideas is vast, in this manuscript we focus on "Topic Models" - a group of techniques arising largely from the computer science literature. As the input to these techniques is textual data (specifically, a collection of text documents) they offer an interesting twist on traditional approaches for understanding the topics and concepts that make up individual publications and, in turn, estimating document similarities and clustering. As discussed below, these techniques are certainly not without their flaws but are also well positioned to exploit the rapidly growing body of textual, and perhaps even full text, data.

In this manuscript we develop a robust approach for calculating pair-wise similarities between documents based on state-of-the-art topic modelling techniques. We compute the similarity between researchers which, in turn, allows us to obtain the topical overlap (or proximity) between them. With this text-only approach, we obtain a continuous knowledge domain space from which we can cluster and delineate topics as narrowly as desired, estimate interdisciplinarity, and observe the evolution and direction of research.

Background

Topic models are a family of statistical algorithms designed to discover the underlying structure of textual data. They rely on the existence of a latent subspace of topics, to which each word is associated with a certain probability. They are a subclass of dimensionality reduction algorithms. In a way, topic models are merely clustering words into reduced sets, ones more easily interpretable by the human eye. This setting allows a two-way transformation: documents, that are originally made up of words, are assigned a number of topics based on the prevalence of the words used. On the other hand, words are clustered into topics, giving the latter a semantic structure. These techniques allow for an automated probabilistic classification of large corpora of text. Originally *machine learning* tools, these approaches have transcended disciplinary boundaries, finding application in scientometrics, as well as a variety of social sciences. The two most widely used and accepted techniques for topic modelling are: Non- negative Matrix Factorization (NMF) (Lee & Seung, 1999) that decomposes into two matrices with non-negative elements, and the better known Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), a refinement of Probabilistic LSA which performed better for document decomposition.

Recent Neural Network based approaches for discovering latent semantic structure in text have emerged from the computer science literature and are quickly growing in popularity. In this line of research topic modelling techniques have evolved to capturing the semantics of individual units of text - from words to morphemes. The algorithms generate embeddings (vectors) for each unit of coherence, in a latent space that has no intrinsic *topic* associated to it. Standalone semantic units are the corpora both in textual data (input) and output of the model. Notably Word2Vec, developed by (Mikolov, Chen, Corrado, & Dean, 2013), is an algorithm based on neural networks that generates word embeddings (word vectors) based on word context. It is substantially different from the previous dimensionality reduction approaches, in that it doesn't try to reduce the (latent) space based on document co-occurrence (bag of words), but rather tries to extract the semantics (meaning) associated to each word based on its surrounding text. More importantly, it overcomes the instability present in the optimization of dimensionality reduction algorithms due to initialization differences. We explore how we use

Doc2Vec to capture a researcher's closest substitutes (those whose output is the most similar) and in doing so, how we overcome the main drawbacks of prior art in topic modelling. We provide a model that can be applied to multiple tasks and that is robust to retraining.

Topic modelling with Doc2Vec

Doc2Vec (Lee & Seung, 1999) is an extension to Word2Vec, working with paragraphs as semantic units, rather than words. It generates document embeddings (vectors) in a latent space in such a way that semantically similar documents are clustered. We use this to our advantage, concatenating text from different documents into one researcher-document. Hence, we construct *researcher embeddings* - each researcher is assigned a vector derived from training the textual data - whose closeness can be evaluated with standard distance metrics. The input text is a combination from PubMed metadata and Author-ity (Torvik & Smalheiser, 2009), which provides a disambiguated set of Author-publication pairs with free text (Titles and Abstracts) and a controlled vocabulary of medical terms, known as MeSH.

Doc2Vec provides consistent results across different sizes of the latent space, with only marginally decreasing values of similarity with increasing number of dimensions, as one would expect. This allows for different levels of granularity in the topic models, and thus, a more fine-grained clustering of topics or words. Doc2Vec offers a new approach for topic modelling using document embeddings. We apply it publication metadata, grouping a researcher's production into one document. The model is robust to stochastic initialisations, and does not fix the number of topics *ex ante*, overcoming the most criticized aspects of prior art. Embeddings are non-sparse vectors which place documents in a latent space based on their inherent semantics and groups close substitutes in space. Every trained model directly provides distance metrics between documents (researchers) that can be used further down in the analysis pipeline, be it for clustering, community detection or as a plain similarity measure.

Measuring interdisciplinarity and novelty

Clustering document embeddings around a centroid provides a measure of interdisciplinarity. With researchers as documents, our topic model helps distinguishing those who work on a set of *perfect substitutable* research questions to those between fields (or communities). Both at a micro and macro level, we can identify trans-boundary people whose area of work is between two discipline centroids, and even have a continuous measure of cognitive distance to one pole or the other. With this approach, it is possible to identify generalists and specialists, and track their evolution across time.

Additionally, it is possible to identify emerging areas of interest in the knowledge space, fads or novelties in science. The classifier allows us to ask a new question without the need of retraining the model, so dynamic vectors (using per-year data) are easily obtained. We can re- calculate the vector of a certain researcher by incorporating or omitting information, i.e. we can leave a publication out or include a new one without recalculating the entire model. This is especially useful for studying the dynamics of research interests. One case of special interest is shown in Figure 1. We split the production of scientists in gaps of 5 years, on a model trained using the entire dataset. We then recalculate the embeddings for each researcher in those 5-year windows, and plot the spatial density in each timeframe. Changes in the concentration of researchers are indicators of what domains are growing or dying. This Knowledge-space density can help characterise novelty, trends and hot topics within a field, as well as the dynamics and (in/out) flow of researchers within that domain.

Figure 1 t-SNE transformed knowledge domain density in 5-year windows