Adam Davies

Discourse Communities, Document Similarity, and the Distributional Hypothesis: A Vector Semantics Approach to Word Frequency Analysis

How much can the stereotypical context of a product, organization, or even a person, say about them? What may be reasonably inferred about an individual on the basis of the media they consume, the food they eat, or the social functions they attend? A similar question has been posed by linguists, with respect to word context; and, as it turns out, many believe context to signify a great deal. As John Rupert Firth famously said, “You shall know a word by the company it keeps.” This is one way of phrasing the *distributional similarity hypothesis*: by examining the context in which a word occurs, we may learn a great deal about its meaning – in particular, by comparing it to other words with similar contexts.

However, one might approach the question of distributional semantics from the opposite direction: to what extent do the words which frequently occur in a given context define the context itself? This perspective receives much less attention, but is nearly equivalent. The question with which this project is primarily concerned is: *to what extent may we understand a context by virtue of the words which are found within it?* This question may be posed in terms of *discourse communities* – particularly those defined by a given academic discipline – and the many contextual features which they influence, or even define. As noted above, a thorough understanding of a given linguistic context is, in more than one way, dependent upon one’s familiarity with the idiosyncratic features of its lexicon.

The research conducted as a part of this project suggests that the relationship between discourse communities and more narrow forms of context – in this case, individual textual documents – is quite analogous to the relationship between those contexts and the words within them. To study the former relationship, a large number of documents across four discourse domains were statistically analyzed, using vector-based measures of similarity, in the attempt to quantify the strength of the distributional hypothesis across areas of discourse. These four discourses are law, mathematics, medicine, and philosophy. Each of these disciplines have fostered its own respective discourse community with a highly specialized lexicon; as such, examining the relationships among the documents which may be found in any one of these discourses, and comparing them to documents in the others, should be especially informative. To

study these relationships, we employed the *cosine similarity measure* among documents in the same discourse, in comparison to the same between documents in distinct discourses. The distributional similarity hypothesis, interpreted in this context, predicts a substantially higher cosine similarity score among the former than the latter.

The materials which were used to carry out this study were made available by the researchers from English-Corpora.org, which assembled each of the four lists of 100 Wikipedia articles closely related to one of the areas of discourse under investigation, and compiled this data in a workable format. Total, the size of all four corpora amounts to over 1 million words (see Table 1), which proved more than enough to observe a large contextual discourse effect.

Law	Mathematics	Medicine	Philosophy
304505 words	277435 words	295577 words	323159 words

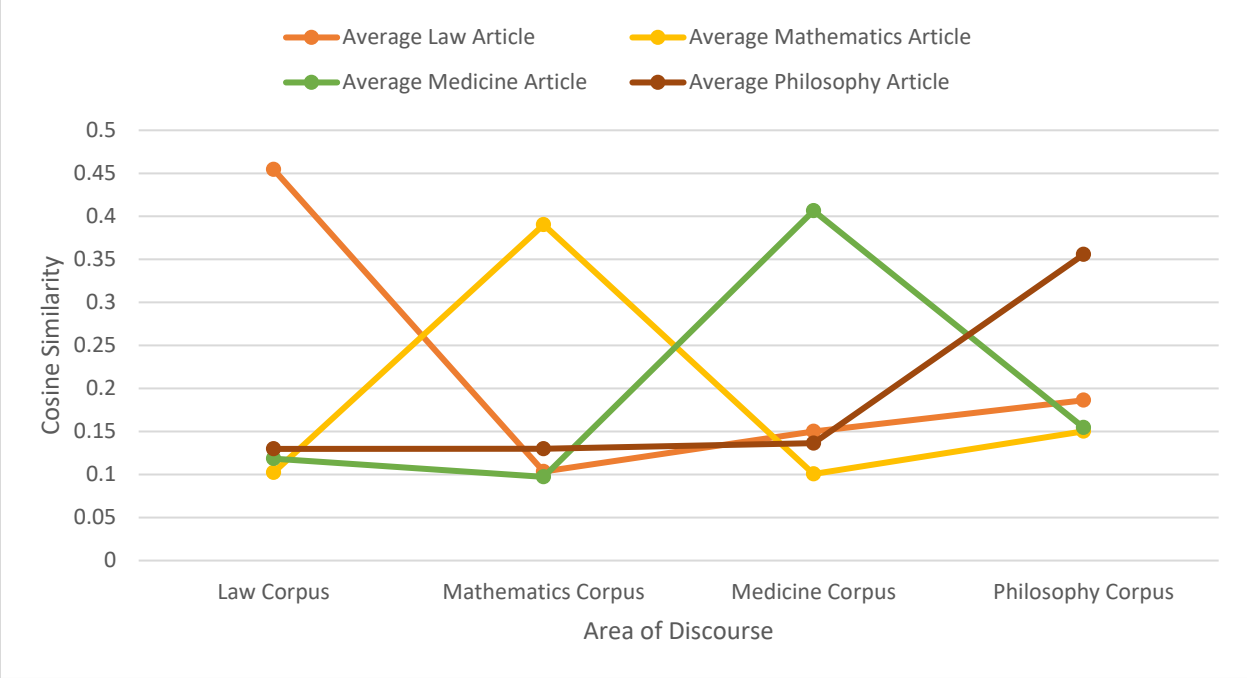
Table 1: the sizes of each of the corpora

As noted above, the relationships between broad domains and the articles among them were assessed on the basis of cosine similarity, which is a high-dimensional analogue of "measuring the angle" between two vectors. The entries of each vector were formed by summing the number of occurrences of a given word throughout an entire document, and subsequently compared to other vectors in the same discourse. As seen in Equation 1, the cosine similarity of two vectors is their dot product, scaled by the norms of both vectors; equivalently, it may be regarded as the *normalized* dot product of both vectors, which ranges strictly from 0 to 1 (inclusive). In this study, cosine similarity cannot be negative, as there are no negative occurrences of words in a document (which could negatively conflict with positive occurrences, yielding a negative cosine measure), meaning that the maximally distinct case for two vectors is when their cosine similarity is 0, in which case they are *orthogonal*. On the other hand, a cosine score of 1 means that the two vectors are, in geometrical terms, "pointing in precisely the same direction", and are simply scaled versions of each other.

$$cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_i A_i B_i}{\sqrt{\sum_i A_i^2} \sqrt{\sum_i B_i^2}}$$

Equation 1: the formula for finding the cosine similarity of two vectors (A and B)

Figure 1: results



	Law Corpus	Mathematics Corpus	Medicine Corpus	Philosophy Corpus
Average Law Article	0.454422622	0.103453593	0.150079821	0.186336325
Average Mathematics Article	0.102346094	0.390277887	0.100592438	0.150185002
Average Medicine Article	0.11839338	0.097248011	0.406489214	0.154505333
Average Philosophy Article	0.129740701	0.129834532	0.136400524	0.355653262

As suggested by the results in Figure 1, the research results were decisive. For each type of article, the cosine similarity observed between members of its own discourse was at least twice as high as it was among members of external discourses. Thus, the distributional hypothesis appears to be quite tenable from the perspective of discourse context, as compared to the more traditional point-of-reference, individual words. If a given document is at all consistent with the features of its own discourse, it will most often exhibit readily identifiable features which distinguish it from documents external to its community.