AI & SOCIETY (2020) 35:135–146
https://doi.org/10.1007/s00146-018-0851-7

ORIGINAL ARTICLE

AI recognition of differences among book-length texts

Stephen J. DeCanio[1]

Received: 22 February 2018 / Accepted: 3 June 2018 / Published online: 19 June 2018

Abstract

Can an Artificial Intelligence make distinctions among major works of politics, philosophy, and fiction without human assistance? In this paper, latent semantic analysis (LSA) is used to find patterns in a relatively small sample of notable works archived by Project Gutenberg. It is shown that an LSA-equipped AI can distinguish quite sharply between fiction and non-fiction works, and can detect some differences between political philosophy and history, and between conventional fiction and fantasy/science fiction. It is conjectured that this capability is a step in the direction of "M-comprehension" (or "machine comprehension") by AIs.

Keywords Artificial Intelligence · Latent semantic analysis · Natural language processing · Textual analysis

1 Introduction

What does it mean to understand text? The problem of meaning has occupied philosophers from the beginning of systematic speculative thinking (Landauer 2011). Any AI attempting to extract meaning from written texts would at a minimum need to distinguish between different kinds of works. Recognition of the genre, style, and content of texts constitutes a first step. There is already a sizeable literature on "natural language processing" (NLP), or automated extraction of information from text. Others have surveyed this literature (Foltz 1998; Gomaa and Fahmy 2013; Mikolov et al. 2013a, b; Shiffrin and Börner 2004), and no effort will be made here to catalogue every approach. Examples include "probabilistic topic models" such as latent Dirichlet allocation (LDA) that use Bayesian methods to extract both topics and structural relationships from underlying bodies of text (Steyvers and Griffiths 2011; Blei 2012), and neural networks that can be applied to a variety of NLP tasks (Collobert and Weston 2008). Modern "stylometry", the study of linguistic style, applies statistical methods to questions of authorship attribution.[1]

Perhaps the simplest and one of the best-established NLP system is "latent semantic analysis" (LSA). LSA can be implemented with off-the-shelf software and ordinary computational hardware.[2] It is scalable, and has a record of success in autonomous learning, essay grading, diagnosing schizophrenia, and information retrieval.[3] LSA has been used for similarity analysis of the titles of scientific papers, to show a decline in international cooperation and research productivity after 1914 (Iaria et al. 2017). Computer systems with these capabilities are far from "understanding" or "comprehending" the texts they analyze, but they are mimicking many human capabilities. As for the potential of LSA, one of its leading proponents has asked:

> Suppose we have available a corpus of data approximating the mass of intrinsic and extrinsic language-relevant experience that a human encounters, a computer with power that could match that of the human brain, and a sufficiently clever learning algorithm and data storage method. Could it learn the meanings of all the words to any language it was given? (Landauer 2011, p. 4).

In the present paper, terms such as "M-comprehension" or "M-understanding" will be used to indicate the capabilities of actually functioning computers. There is no doubt

✉ Stephen J. DeCanio
decanio@econ.ucsb.edu

[1] Economics, Emeritus, University of California, Santa Barbara, USA

[1] See the Wikipedia article "Stylometry" (2018) and the references therein.

[2] All of the analysis in this paper was carried out running Mathematica (Wolfram Research Inc., 2017) on an ordinary PC.

[3] A convenient reference point for LSA is Landauer et al. (2011). The first essay in this complication, Landauer (2011), gives a sample of the successful applications of LSA with associated references.

that strong pattern-recognition capacity can be achieved with existing hardware and software, but how finely can an AI using LSA identify differences and similarities between book-length texts? I propose to test whether an LSA-equipped AI can make distinctions among significant works of political philosophy, history, and fiction. A modest number of texts were analyzed—a corpus of 100 major works drawn from the list of frequently downloaded books compiled by Project Gutenberg.[4]

In LSA, words and documents are coded as a matrix (a row for each word, a column for each document) which is then condensed to a "semantic space" or "concept space" of lower dimensionality. The element t_{ij} of the raw word-document matrix equals the number of times word i appears in document j. Entries of the term-document matrix are then weighted to give a relatively high weight to elements that occur frequently in some but not all of the documents and relatively low weight to words that appear frequently throughout the corpus. Again, there are different ways this can be done, but experience has shown that "log-entropy" weighting performs well.[5] The weighted word-document matrix will be denoted by $\mathbf{A}$, an $m \times n$ matrix with m rows (one for each word) and n columns (one for each document). This particular weighting scheme is discussed in Martin and Berry 2011, pp. 37–39, citing Dumais (1991), Salton and Buckley (1991), Letsche and Berry (1997), and Berry and Browne (2005). Unlike most of the applications of LSA found in the literature, the "documents" in the present paper consist of entire books, not just paragraphs or short passages.

It might be possible for an AI to use the unreduced word-document matrix to identify similarities between different documents in the corpus. Similarities can be defined in a variety of ways. The simplest is to calculate the cosine between the column vectors representing any pair of texts in the weighted term-document matrix $\mathbf{A}$. The cosine of the angle between two vectors $\mathbf{s_1}$ and $\mathbf{s_2}$ in a vector space is given by $\frac{\mathbf{s}_1 \cdot \mathbf{s}_2}{\|\mathbf{s}_1\| \, \|\mathbf{s}_2\|}$, where the numerator is the dot product of the two vectors and $\|\mathbf{s_i}\|$ is the ordinary Euclidean norm (length) of $\mathbf{s_i}$. A pictorial representation of the 100×100 table of these cosines is given in Fig. 1, with each of the 10,000 squares at locations (j_1,j_2) in the Figure representing the cosine between document j_1 and document j_2.

Figure 1 is an easier way of seeing the patterns in the cosine pairs than a 100×100 table of numbers printed in a tiny font. The shading in the table goes from lighter (smaller cosines) to darker (larger cosines). Quite clearly, the squares down the main diagonal are darkest, because the cosine of a vector with itself is 1. The average degree of similarity (average cosine) in the whole matrix is 0.207 with standard deviation 0.106. It is worth noting that AIs have been quite successful in image recognition (Li et al. 2012; Strang 2016), so there is no loss of interpretability in presenting the results of cosine calculations in this way. However, similarities and differences across books as shown by the cosines in Fig. 1 are not particularly strong.

2 Measuring similarities by singular value decomposition

A better approach, the one employed in LSA, is first to factor the $\mathbf{A}$ matrix by singular value decomposition (SVD) (Martin and Berry 2011; Berry and Browne 2005). SVD splits up the $\mathbf{A}$ matrix in a way that makes it easier to identify the concepts or genres that underlie the corpus. Many introductions to SVD are given in the literature, so only an outline of the mathematics is given in Appendix 2. Standard software packages like Mathematica and MATLAB include built-in routines to carry out the SVD calculations. The key step in identifying the strongest similarities involves reducing the information in $\mathbf{A}$ to a "concept space" of markedly lower dimension. Even with as few as 2 or 3 dimensions in the concept space, unsupervised computations clearly distinguish the main types of text in the corpus.

The crucial equation in SVD is $\mathbf{A}_k = \mathbf{U}_k \, \mathbf{\Sigma}_k \, \mathbf{V}_k^{\mathbf{T}}$ (see Appendix 2). Here k is the dimension of the concept space, $\mathbf{A}_k$ is an $m \times n$ matrix, $\mathbf{U}_k$ is an $m \times k$ matrix, $\mathbf{\Sigma}_k$ is a $k \times k$ diagonal matrix (all off-diagonal elements are zeros), and $\mathbf{V}_k^{\mathbf{T}}$ is a $k \times n$ matrix. The diagonal elements of $\mathbf{\Sigma}_k$ are the "singular values" ranked from largest to smallest. Essentially, SVD "diagonalizes" the $\mathbf{A}$ matrix and finds the "right" bases for its associated fundamental subspaces (Strang 2016). Following Martin and Berry (2011), the column vectors of $\mathbf{V}_k^{\mathbf{T}}$, scaled by the corresponding singular values, are the "document vectors". Here they will be denoted by $\mathbf{v}_j$, and they are the vectors that will be analyzed for similarity in the concept space. With the 100 texts considered here, j ranges from 1 to 100.

Similarity will be measured as the cosine between document vectors in the reduced space. However, it should first be noted that the lengths of each of the $\mathbf{v}_j$ vectors, as well as the first elements of those vectors, are highly dependent on

[4] Project Gutenberg (2018) offers electronic versions of books that are freely available to the public. The texts can be used in any appropriate way provided Project Gutenberg is acknowledged, which acknowledgment is gratefully given here. A full list of the books in the corpus used for this analysis is given in Appendix 1.

[5] The Log-Entropy weighting of element a_{ij} in the matrix $\mathbf{A}$ is defined by $a_{ij} = \text{local}(i,j) \times \text{global}(i)$. Here, $\text{local}(i,j) = \log_{10}(1 + t_{ij;})$. Entropy is defined as $-\sum_i \frac{p_{ij} \, log_2(p_{ij})}{log_2 \ n}$, with $p_{ij} = \frac{t_{ij}}{g_i}$, t_{ij} the frequency of word i in document $j;$ n is the total number of documents; and g_i is the number of times the word i appears in the entire corpus. Then $\text{global}(i) = 1 - \mathit{Entropy}$. Note that if $t_{ij} = 0$, then $\text{local}(i,j) = 0$. This guarantees that the $\mathbf{A}$ matrix sill be "sparse", (i.e., will contain mostly zeros), which speeds up computations for the LSA.

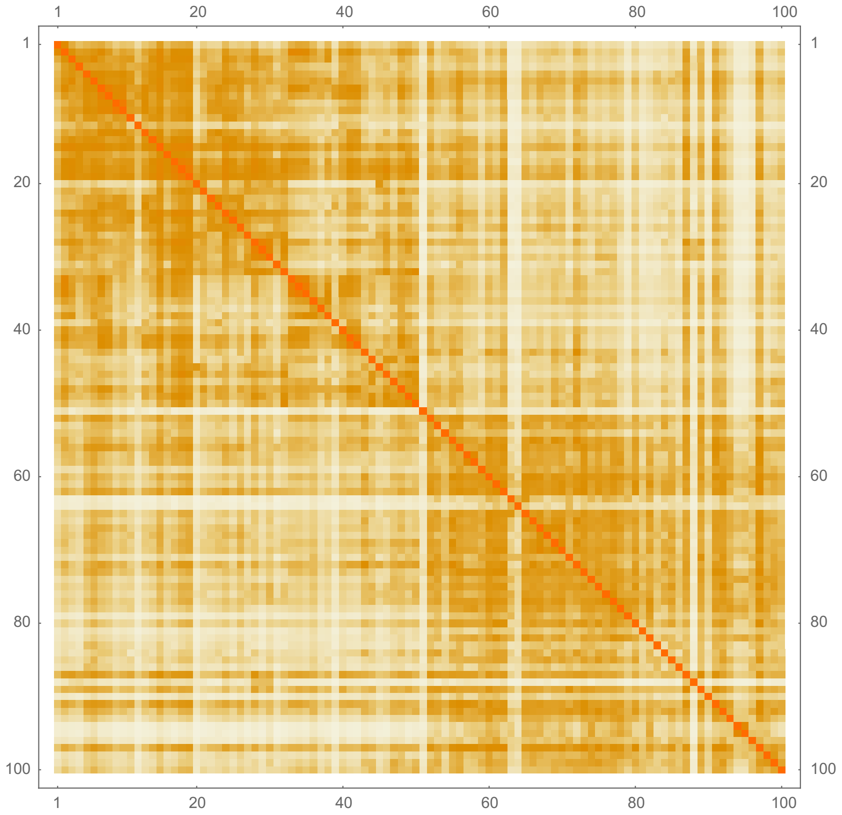


Fig. 1 Cosines between document pairs (vectors) in the **A** matrix; Darker shade indicates greater similarity

the sheer length of the texts indexed by j. The correlation between the string lengths of the texts and lengths of the $\mathbf{v}_j$ is 0.808, and the correlation between the absolute values of the first elements of the $\mathbf{v}_j$ and the string lengths is 0.885. It is clear that the cosine similarity between two dissimilar vectors can be dominated by the vectors' lengths if one component is much larger than the others. For example, the cosine between $\{30, -2, 1\}$ and $\{20, 1, 1\}$ is 0.9931, and the cosine between $\{30, 2, 1\}$ and $\{20, 1, 1\}$ is 0.9997. If the large first components are ignored, the cosine between $\{-2, 1\}$ and $\{1, 1\}$ is -0.316, while the cosine between $\{2, 1\}$ and $\{1, 1\}$ is 0.949. The first component of these 3-component vectors makes it seem that the two vectors are close regardless of whether the second component is 2 or -2, whereas the vectors made up of only the second and third components are highly dissimilar. SVD is related to principal component analysis (Shirota and Chakraborty 2015; Shlens 2014), and quite obviously the largest variation among the document vectors will be in the direction of the length component.[6] The first weighted row vector in $\mathbf{V}_k^{\mathbf{T}}$ represents primarily the length of the texts. Therefore, if the weighted $\mathbf{v}_j$'s, excluding their first components, are projected onto a lower-dimensional subspace, it is possible to visualize similarities or differences of the most important concepts other than length. Figure 2 shows the cosines between pairs of vectors made up of the second and third components of the columns of $\mathbf{V}_k^{\mathbf{T}}$ for $k = 3$.

Although the complexity of the concepts contained in the full corpus is not captured by this reduced space, a sharp discrimination among the 100 books is possible. The non-fiction works are distinct from the novels. In Fig. 2, the color scheme goes from "hot" (red) to "cold" (blue) as the cosine decreases from + 1 towards − 1. The books have been numbered from 1 to 100, with the first fifty being the works of political philosophy, economics, and history and the bottom fifty being works of fiction. This ordering was done to facilitate exposition and to make the patterns clear to human readers, but it would not be necessary for an AI's M-comprehension of the texts. The AI could easily do its own ordering based on the cosine similarity measures. Book numbers are shown in Appendix 1, and points on the axes in Fig. 2 correspond to the book numbers. The numbered tick marks are in intervals of 20 from 1 to 100.

In Fig. 2 the books split cleanly into the fiction and non-fiction groups, with each block having high within-group similarity and low out-of-group similarity. The political/philosophical/historical books are the red–orange rectangle in the upper left, while the novels are in the red–orange block in the lower right. The blue off-diagonal blocks indicate that the cosines between books in the two different groups are negative. The contrast between within-group similarity and

[6] This is pointed out by Hu et al. (2011) citing Buckley et al. (1996), and by Bhagwant (2011).

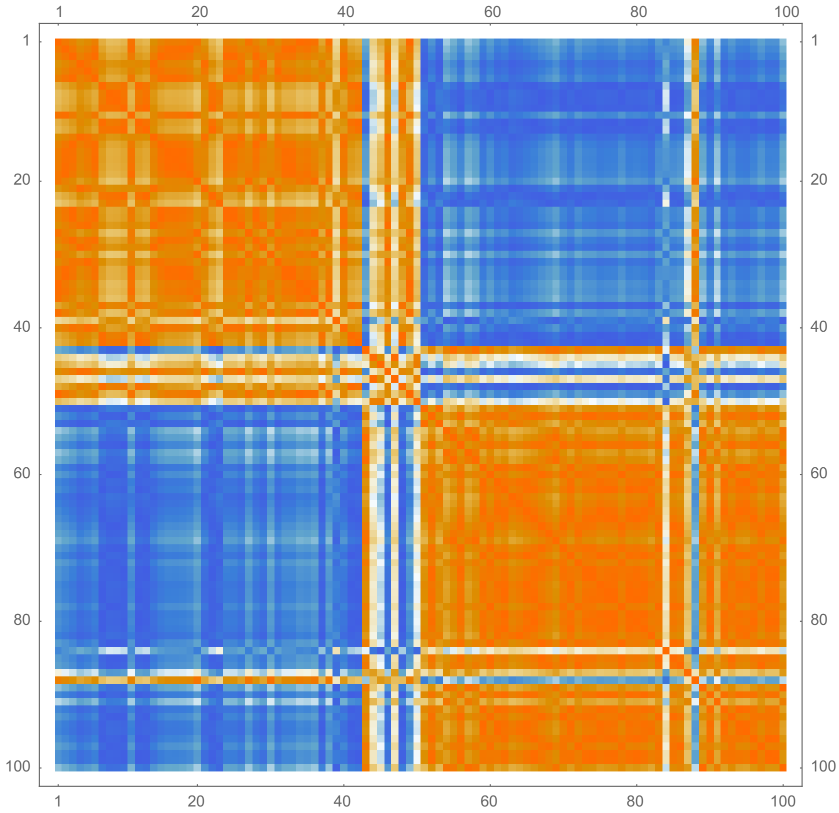


Fig. 2 Graphic representation of cosines between column vectors in $\mathbf{V}_k^{\mathrm{T}}$, dropping first component of each vector, $k = 3$

Table 1 Cosine averages and standard deviations for blocks of books

	All books	Non-fiction/non-Fiction	Non-Fiction/fiction	Fiction/fiction
Figure 2, SVD with $k = 3$ (length components of the v_j omitted)				
Mean cosine	0.061	0.758	**− 0.685**	0.855
Standard deviation	0.851	0.394	0.445	0.339
Figure 1, raw cosines				
Mean cosine	0.207	0.258	0.165	0.239
Standard deviation	0.106	0.129	0.050	0.126

out-of-group similarity can be summarized by the averages of the cosines in the different blocks of Fig. 2. Table 1 also contrasts the strength of the fiction/non-fiction distinction found with SVD (Fig. 2) to the relatively mild differentiation in the blocks shown in the depiction of raw cosines (Fig. 1).

The difference between the fiction and non-fiction books could hardly be clearer. The within-group cosines are almost always close to 1, while the across-group cosines are almost all negative. Even the exceptions are informative. Among the non-fiction works, Carlyle's *The French Revolution* (book #43) has a negative cosine when compared to the other non-fiction books. Indeed, Carlyle's style is novelistic (Hindley 2009). At the same time, among the works of fiction, Swift's *A Modest Proposal* (book #88) is an outlier. But of course, the "modest proposal" was a vicious satire, suggesting that the problem of poverty in Ireland could be solved by cannibalizing the island's 1-year-old children. The horror of butchering and eating babies presented as though it were a serious policy proposal. In other words, *A Modest Proposal* was meant to read *as if it were* non-fiction.

More instances of the ability of the three-dimensional reduced space to pick out unusual books can be seen by expanding the non-fiction and fiction blocks, as is done in Figs. 3 and 4. Consider Fig. 3, the graphic representation of the cosines between the non-fiction books (numbered tick marks are in intervals of 10). In addition to the Carlyle history (#43) already pointed out, it seems that the more modern historical works show somewhat weaker similarity to the ancient historians and the political philosophers. The vectors associated with Grant's *Memoirs*, Churchill's *River*

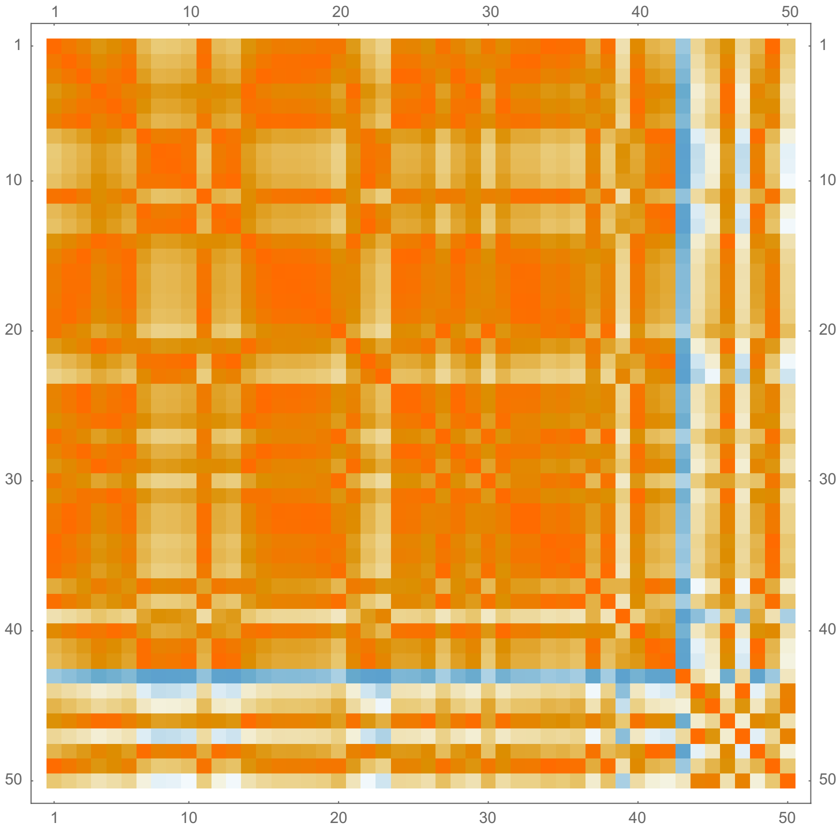


Fig. 3 Graphical representation of cosines for pairs of non-fiction works

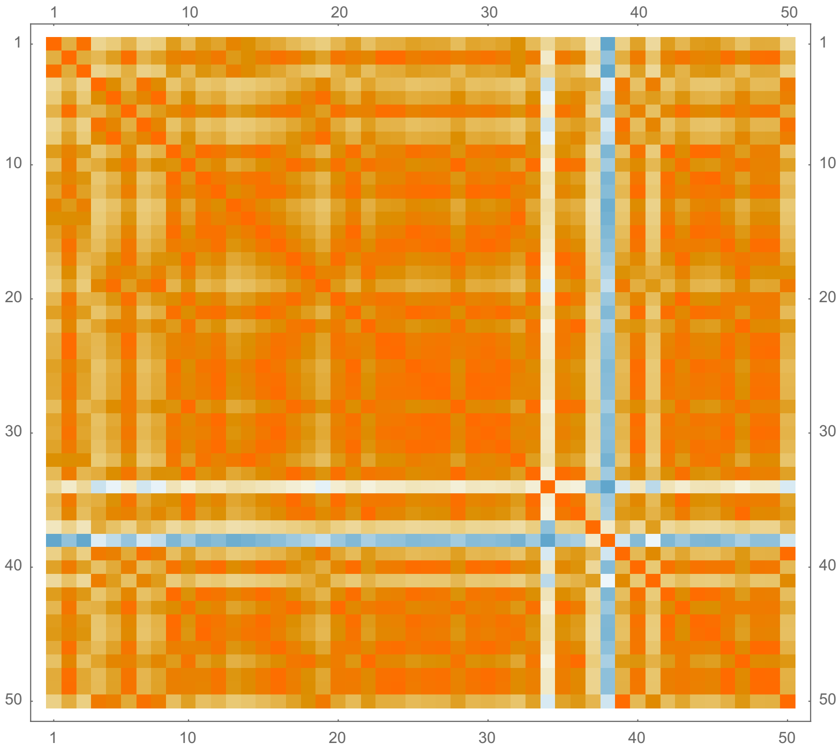
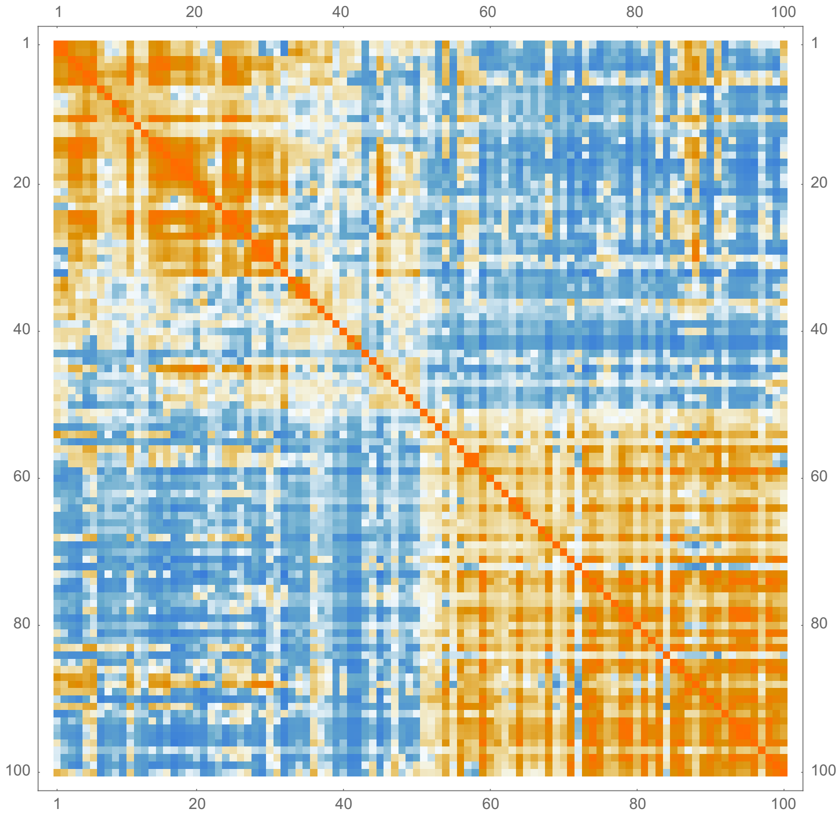


Fig. 4 Graphical representation of cosines for pairs of novels (add 50 to axis numbering to match the numbering of works in Appendix 1)

Fig. 5 Cosine plot for all 100 works, $k=50$



War, and the March and Beamish's *History of the World War* have more than a few negative cosines, but without any dominant pattern.

Figure 4 shows the cosines between novels. (Note that the automatically generated tick marks of Fig. 4 need to have 50 added to match the numbering of the books in Appendix 1, and are numbered in intervals of 10 from 1 to 50). In addition to the anomalous *Modest Proposal* (#88) it is also clear that Joyce's *Ulysses* (#84) is an outlier. The average cosine between *Ulysses* and the other novels is 0.158, while almost all the other cosines are greater than 0.8. (The third lowest fiction average cosine is Swift's *Gulliver's Travels* at 0.627). Of course, *Ulysses* is quite different from typical works of fiction because of its "stream of consciousness" structure (or lack thereof).

Not much additional discrimination among the works shows up as the number of dimensions of the concept space is increased to 4 or 5. However, if the document vectors are projected into higher-dimensional spaces, finer distinctions among the different works are possible. Instead of $k=3$, consider a SVD with $k=50$. Again, ignore the first component of each of the $\mathbf{v}_j$ vectors to reduce the influence of length on the closeness measure. The 100×100 plot of this cosine matrix is shown in Fig. 5.

Once again, the similarity measures fall into blocks. Books within the fiction block show positive similarity (orange-colored squares). As before, the non-fiction and fiction books have negative (bluish-colored squares) cosines, with a few exceptions. Both of Swift's satires are similar to many of the non-fiction books.

More interestingly, within the non-fiction block the histories show less similarity to the books that are pure philosophy or political theory. This is illustrated in Fig. 6.

The histories begin with Titus Livius I (#33) and continue through March and Beamish's *History of the World War* (#50). The frequency of negative cosines with the pure political theory and philosophy works is clearly greater for the histories. It also seems that the ancient histories (Titus Livius (#33) through Gibbon II (#42)) show a degree of similarity to each other, as do the "modern" histories after Carlyle (Grant (#44) through March and Beamish (#50)), although these similarities are not particularly strong. The strongest similarities in Fig. 6, however, are the ones between the works of political theory and philosophy (considered as a group). The only strong negative cosines in the upper left corner are between Machiavelli (#1 and #2) and the "modern" economists—Ricardo (#29), Jevons (#30), Veblen (#31) and Keynes (#32)!

Fig. 6 Cosine plot for non-fiction works, $k=50$

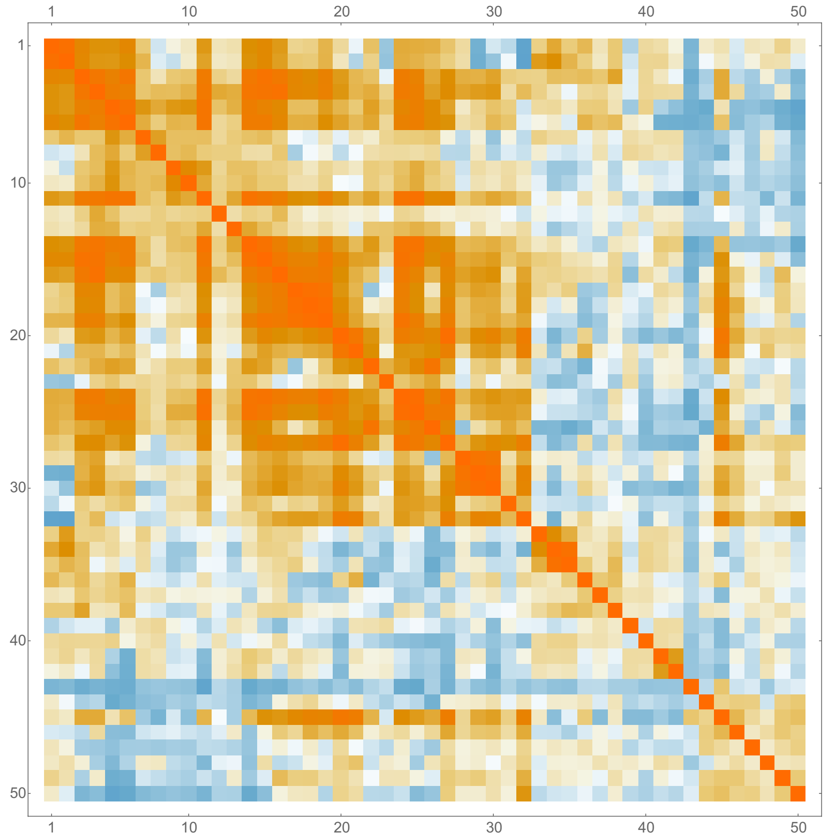
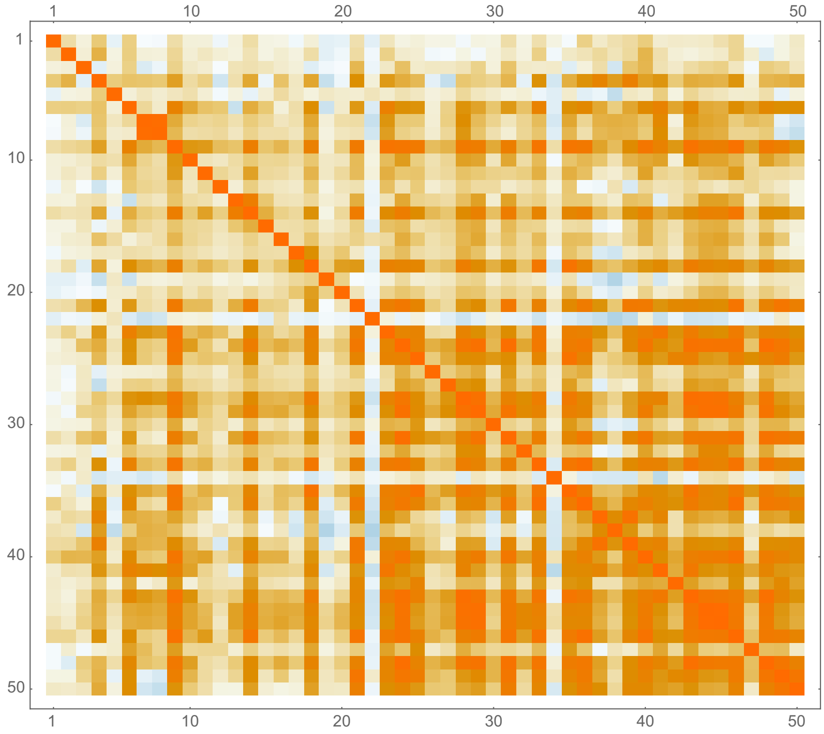


Fig. 7 Cosine plot of novels, $k=50$



The fiction works also show finer distinctions than in the 3-dimensional case: The books were arranged roughly from the least recent to the more recent, except that the fantasy/science fiction novels are grouped together at the bottom of the list. Figure 7 shows that, in general, the more recent the novel, the greater its similarity with other novels in the fiction group.

In Fig. 7, the coloration becomes darker (greater similarity) moving from the upper left corner (the earliest works) down to the lower right corner (most recent works). But it is also clear that the fantasy and science fiction books can be picked out—these are the works after *Ulysses* (#84; axis point 34 after subtracting 50).

What is seen in Figs. 5, 6 and 7 is that the SVD model can identify sub-categories within the two larger non-fiction and fiction groups. This is because the vectors projected into the 50-dimensional subspace contain more information about the books in the corpus than when only three dimensions are retained. However, the patterns of the 50-dimensional SVD are essentially unchanged if the document vectors are projected into the 100-dimensional space that is the maximum obtainable with this corpus of 100 works. Increasing the number of dimensions past a certain point provides no improvement in the resolution of underlying concepts. This is consistent with the observation that in LSA analyses "[t]he number of dimensions retained in LSA is an empirical issue" (Landauer et al. 1998, p. 269).

3 Discussion

To check the robustness of the results, a comparable sample of 100 texts was created with random words selected from the same "dictionary" that encompassed all the words in the books of Appendix 1. Four groups of random "books" were created, with 25 having length 20,000 words, 25 having length 50,000, 25 having length 100,000, and 25 having length 500,000. The mean number of words of the books in the main corpus is 153,666 while the mean number of words for the books in the random corpus is 167,500, so the two sets of texts are roughly comparable in size. The first singular value for the weighted word-document matrix of the random texts is 17 times larger than the second singular value. The last 99 singular values drop off very slowly decreasing by only a factor of about 2 from the second to the 100th singular value. If the first component of the document vectors (the one that is correlated with the length of the documents) is dropped, the resulting projection of vectors composed of the second and third components onto the reduced concept subspace shows a random pattern of cosines. The pictorial representation of this lack of pattern is shown in Fig. 8. The average cosine between vector pairs in Fig. 8 is 0.005, with a standard deviation of 0.711.

Returning to the main results, why is it that the SVD encodes information about the documents more efficiently than the simple comparison of raw cosines? There is no doubt that the reduction in dimensions reduces "noise" present in the sample of documents. The SVD is picking out vectors in the reduced space that correspond to the directions of greatest variance, so elimination of all but the first few components will eliminate some of the noisy elements of the texts. It must be admitted, however, that exactly what the SVD is finding is somewhat mysterious. As one group of researchers put it.

> At this point, there is a great deal of uncertainty about what is being represented in the K-dimensional spaces of LSA. One optimistic possibility is that the K dimensions reflect ontological categories, semantic features, and structural compositions of mental models that would be directly adopted in structural theories of world knowledge representation….[Nevertheless,] [v]ery few researchers would go out on the limb and propose an elegant mapping between the K dimensions of LSA and sophisticated theories of world knowledge. However, most researchers would seriously entertain the possibility of weaker correspondences (Graesser et al. 2000, p. 2, references omitted).

The difficulty would seem to stem from the problem of how the meanings of words are determined, a philosophical question that goes back to Plato (Landauer 2011).

From a broader perspective, one long-term goal of the AI project is teaching an AI to M-comprehend a wide variety of texts. One way of doing this would be to let the AI read voraciously. It is possible that, using the kind of procedure outlined in this paper, an AI would be able to distinguish works of moral and political philosophy from the welter of information that has been digitized, and thereby would be exposed to the full range of literature on human moral systems. Of course, this wide-ranging input would not solve the dilemma posed by the fact that humans do not agree about morality,[7] but it would provide plenty of input for M-thinking about moral issues. This could be a first step in development of M-ethics in AIs (DeCanio 2017). Other approaches to imparting moral values to AIs are possible, too. Guarini (2006) trained a neural network with case-based moral reasoning, and has argued that "aspects of duty can be preserved for machine ethics" (Guarini 2012, p. 434). Wallach and Allen (2009)

[7] The problem of moral disagreement is starkly posed by MacIntyre (1984).

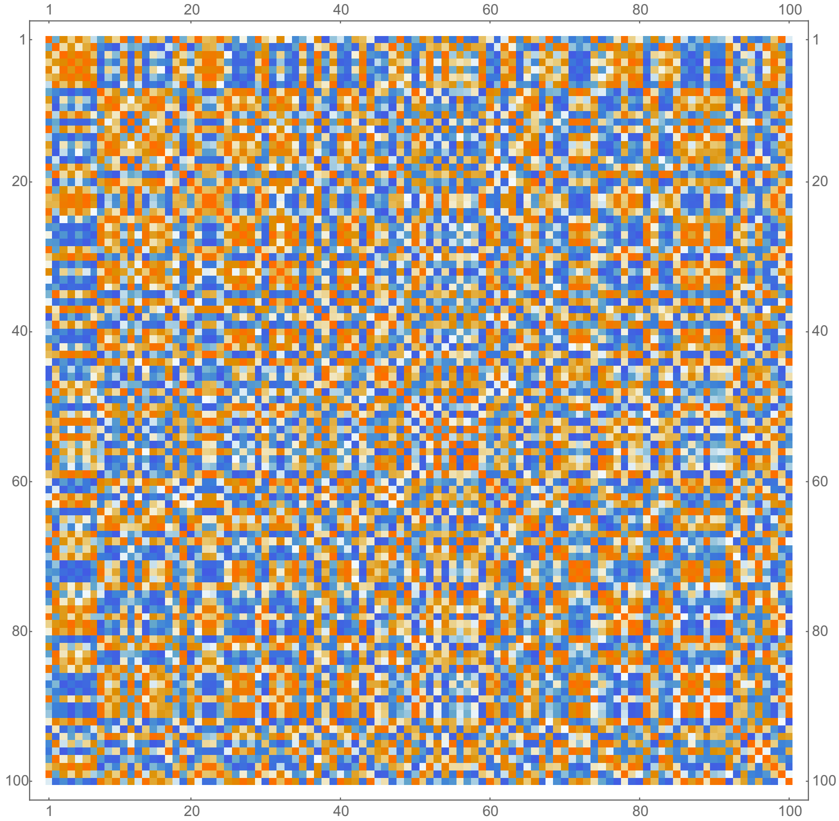


Fig. 8 Cosine plot of 100 random-word documents, length component dropped, $k=3$

and the edited collection by Lin et al. (2014) have explored these issues. Unsupervised computational methods even have been brought to bear on matters of Biblical scholarship (Hu 2012).

Any of the distinctions found among the 100 works analyzed in this paper could have been discovered by an unaided AI. The pictorial grouping into fiction and non-fiction works in Figs. 2 and 5 was done to make it easier for a human reader to see the patterns in the concept space. An AI could have simply used the cosine similarity measure to come up with rankings that would reveal the differences. The AI would have found that the non-fiction books are similar, as are the fiction books, but that each of these groups is dissimilar to the other (with exceptions as noted above). With higher-dimensional concept space, additional distinctions could be drawn. This kind of classification is an initial step towards M-comprehension. It seems plausible that with larger numbers of works included in the database, it would be possible to make finer distinctions. This is a topic for further research.

Is the classification of works into similarity groups equivalent to genuine understanding? With the small amount of data examined here, certainly not. Regardless of how well an AI can classify, highlight, or extract information, the philosophical dilemma posed by the Turing Test remains unsolved. Possession of capabilities, no matter how sophisticated, is not the same as "thinking," but as Turing pointed out, the distinction may be less important than it seems. It should not be underestimated what an unsupervised AI is capable of doing, even with a limited corpus of works. It can tell that there is something "off" about Swift's satires, and that histories do not have the same "feel" as works of pure political philosophy. It can discern evolution of the novel from its earliest examples (*Le Morte d'Arthur* and *Don Quixote*) to the twentieth century, and it can "see" that fantasy/science fiction novels form a similarity group. This is no small achievement for an AI living in an ordinary PC, whose reading list is (so far) only 100 books drawn from Project Gutenberg's archive. There is every reason to believe that the comprehension capabilities of text-interpreting AIs will grow as their literary horizons expand.

Acknowledgements Feedback and suggestions from Alan Sanstad, Tracy DeCanio, and Shirley Pomeranz are gratefully acknowledged. Anne Kirchner provided valuable research assistance. Errors and omissions are those of the author alone.

Appendix 1: Works in the corpus

Political philosophy

- Machiavelli, *The Prince*
- Machiavelli, *Discourses on the First Decade of Titus Livius*
- Aristotle, *Politics*
- Locke, *Second Treatise of Government*
- Aurelius, M*editations of Marcus Aurelius*
- Plato, *The Republic*
- Augustine, *City of God I*
- Augustine, *City of God II*
- Aquinas, *Summa Theologica Part I*
- Aquinas, *Summa Theologica Part I-II*
- Descartes, *A Discourse on Method*
- Hobbes, *Leviathan*
- Leibniz, *Theodicy*
- Hume, *An Enquiry Concerning the Principles of Morals*
- Rousseau, *Social Contract & Discourses*
- Paine, *Common Sense*
- Hamilton, Madison, and Jay, *The Federalist Papers*
- de Toqueville, *Democracy In America I*
- de Toqueville, *Democracy In America II*
- Marx and Engels, *The Communist Manifesto*
- Engels, *The Origin of the Family Private Property and the State*
- Schopenhauer, *The Basis of Morality*
- Schopenhauer, *The World as Will and Idea III*
- Mill, *On Liberty*
- Mill, *Utilitarianism*
- Nietzsche, *Beyond Good and Evil*
- Russell, *Political Ideals*

Economics

- Smith, *The Wealth of Nations*
- Ricardo, *On The Principles of Political Economy*
- Jevons, *Political Economy*
- Veblen, *Theory of the Leisure Class*
- Keynes, *Economic Consequences of the Peace*

History

- Titus Livius, *The Hisstory of Rome I*
- Titus Livius, *The History of Rome II*
- Titus Livius, *The History of Rome III*
- Xenophon, *Anabasis*
- Herodotus, *The History of Herodotus*
- Thucydides, *History of the Peloponessian War*
- Grote, *Historyof Greece*
- Josephus, *The Wars of the Jews*
- Gibbon, *The History of The Decline and Fall I*
- Gibbon, *The History of The Decline and Fall II*
- Carlyle, *The French Revolution*
- Grant, *Personal Memoirs*
- Marx, *The Eighteenth Brumaire of Louis Bonaparte*
- Turner, *The Frontier in American History*
- Churchill, *The River War*
- Wells, *The Outline of History*
- Mahan, *Influence of Sea Power*
- March and Beamish, *History of the World War*

Fiction

- Malory, *Le Morte d'Arthur*
- Cervantes, *Don Quixote*
- Melville, *Moby Dick*
- Voltaire, *Candide*
- Dumas, *The Count of Monte Cristo*
- Hawthorne, *The Scarlet Letter*
- Austen, *Pride and Prejudice*
- Austen, *Emma*
- Dickens, *A Christmas Carol*
- Dickens, *A Tale of Two Cities*
- Alcott, *Little Women*
- Eliot, *Middlemarch*
- Twain, *Adventures of Huckleberry Finn*
- Twain, *The Adventures of Tom Sawyer*
- Stowe, *Uncle Tom's Cabin*
- Flaubert, *Madame Bovary*
- Dostoyevsky, *The Brothers Karamazov*
- Dostoyevsky, *Notes from the Underground*
- Tolstoy, *War and Peace*
- Tolstoy, *Anna Karenina*
- Stevenson, *Treasure Island*
- Hugo, *Les Misérables*
- Brontë, *Wuthering Heights*
- Conrad, *Heart of Darkness*
- Doyle, *Complete Sherlock Holmes*
- Sinclair, *The Jungle*
- Dreiser, *Sister Carrie*
- James, *The Turn of the Screw*
- Wharton, *Ethan Frome*
- Wharton, *The Age of Innocence*
- London, *The Call of the Wild*
- Montgomery, Anne of Green Gables
- Hesse, *Siddhartha*
- Joyce, *Ulysses*

Fantasy/science fiction

- Kafka, *The Trial*
- Rand, *Anthem*
- Swift, *Gulliver's Travels*
- Swift, *A Modest Proposal*
- Defoe, *The Life and Adventures of Robinson Crusoe*
- Irving, *The Legend of Sleepy Hollow*
- Shelley, *Frankenstein*
- Stoker, *Dracula*
- Stevenson, *The Strange Case Of Dr Jekyll and Mr Hyde*
- Carroll, *Alice in Wonderland*
- Carroll, *Through the Looking Glass*
- Baum, *The Wonderful Wizard of Oz*
- Collins, *The Woman in White*
- Wells, *The Time Machine*
- Wells, *The War of the Worlds*
- Burroughs, *A Princess of Mars*

A few things should be pointed out about the corpus. No evaluation of literary qualities was involved in creating the corpus. In several cases, an author was represented not by his or her most noted work, because the list of text was limited by what is in the Project Gutenberg archive. Pre-processing involved elimination of punctuation, capitalization, and "stopwords" such as *a*, *and*, *the*, etc. The words were not "stemmed" (that is, eliminating all but the root of plurals, words ending in -ed, -ing, and so forth). Mathematica can perform its proprietary version of "Porter stemming", (Porter 1980) but the criteria are not entirely transparent and stemming is not without problems (Wikipedia 2017). Given the lengths of the documents being analyzed, stemming is perhaps less important than it might be with shorter documents. Introductions by editors and/or translators other than the authors of the works were removed. The texts contain various kinds of "noise" such as footnote demarcations, chapter numberings, spelling variations, etc. Translated works will necessarily embody some aspect of the translator's "voice" in addition to the author's, but that should not be too great a problem at the level of analysis being carried out in the paper.

Appendix 2: The basic mathematics of singular value decomposition

The notation used here is fairly standard, and the derivation largely follows Martin and Berry (2011). It is assumed that $m >> n$ in the $\mathbf{A}$ matrix because there are many more words than books in the corpus. (In the analysis here, $m = 130{,}342$ and $n = 100$.)

In matrix notation, SVD factors the $\mathbf{A}$ matrix as follows:

$$\mathbf{A} = \mathbf{U}\,\boldsymbol{\Sigma}\,\mathbf{V}^{\mathbf{T}}, \tag{1}$$

where $\mathbf{A}$ is the word-document matrix, $\mathbf{U}$ is the orthonormal matrix made up of the eigenvectors of $\mathbf{A}\ \mathbf{A}^{\mathbf{T}}$ (with $\mathbf{A}^{\mathbf{T}}$ the standard matrix notation for the transpose of $\mathbf{A}$), $\boldsymbol{\Sigma}$ is the (diagonal) matrix of singular values, and $\mathbf{V}$ consists of the orthonormal eigenvectors of $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ (Strang 2016). If there are m words and n documents in the corpus, $\mathbf{U}$ will be a $m \times n$ matrix, $\boldsymbol{\Sigma}$ will be an $n \times n$ diagonal matrix (all off-diagonal elements zero), and $\mathbf{V}$ will be a $n \times n$ matrix. The singular values associated with the 100-item corpus are displayed in Fig. 9, with these singular values ordered from largest to smallest. The $\mathbf{A}$ matrix can be approximated without too much loss of information using SVD. The reduction of dimensions creates a "concept space" in which the most important semantic and structural features are preserved. Projecting the individual texts into this low-dimensional concept space enables clustering of texts into groups that are conceptually or stylistically similar. In matrix terms,

$$\mathbf{A}_k = \mathbf{U}_k\,\boldsymbol{\Sigma}_k\,\mathbf{V}_k^{\mathbf{T}}, \tag{2}$$

where k is the parameter of the reduced space chosen by the analyst. The reduced space is obtained by setting all columns of $\mathbf{U}$ greater than k equal to zero, setting all singular values smaller than the kth equal to zero, and setting all columns of $\mathbf{V}$ greater than k equal to zero. This decomposition is shown pictorially in Fig. 10.

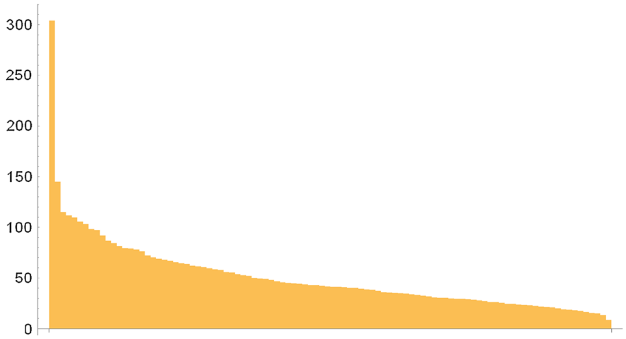


Fig. 9 Singular values 1–100, in descending order of magnitude

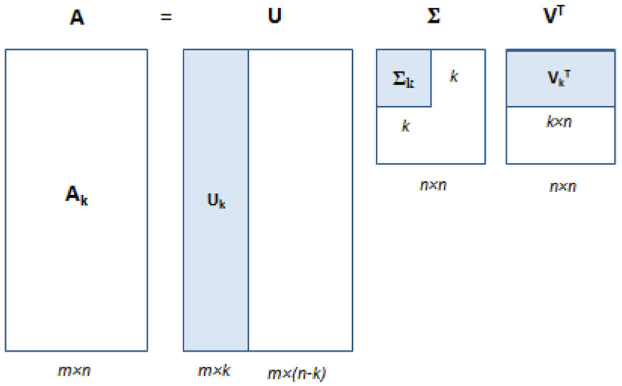


Fig. 10 SVD of matrix A

The $\mathbf{A_k}$ matrix is an approximation of the $\mathbf{A}$ matrix, but one that is made up out of smaller factors than the factors of $\mathbf{A}$ shown in Eq. (1). With $k=3$, for example, $\mathbf{\Sigma_k}$ will be a 3×3 matrix consisting of the three largest singular values on the diagonal and zeros off the diagonal, and the n columns of $\mathbf{V}_k^{\mathbf{T}}$ will be images of the document vectors in the (reduced) three-dimensional concept space. The columns of $\mathbf{V}_k^{\mathbf{T}}$ scaled by the corresponding singular values in Σ_k constitute the "document vectors" and are denoted as $\mathbf{v}_j$, with j ranging from 1 through n.

References

Berry MW, Browne M (2005) Understanding search engines: mathematical modeling and text retrieval. Society for Industrial and Applied Mathematics, Philadelphia

Bhagwant (2011) Latent semantic analysis (LSA) tutorial. https://technowiki.wordpress.com/2011/08/27/latent-semantic-analysis-lsa-tutorial/. Accessed 7 June 2018

Blei DM (2012) Probabilistic topic models. Commun ACM 55(4):77–84

Buckley C, Singhal A, Mitra M, Salton G (1996) New retrieval approaches using SMART: TREC 4. In: Harman DK (ed) The fourth Text REtrieval Conference (TREC-4). US Department of Commerce (NIST Special Publication 500–236)

Collobert R, Weston J (2008) A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th international conference on machine learning. Helsinki, Finland

DeCanio SJ (2017) Games between humans and AIs. AI & Society https://doi.org/10.1007%2Fs00146-017-0732-5

Dumais ST (1991) Improving the retrieval of information from external sources. Behav Res Methods Instrum Comput 23(2):229–236

Foltz PW (1998) Quantitative approaches to semantic knowledge representations. Discourse Process 25(2–3):127–130

Gomaa WH, Fahmy AA (2013) A survey of text similarity approaches. Int J Comput Appl (0975–8887) 68(13):13–18

Graesser A, Karnavat A, Pomeroy V, Wiemer-Hastings K, The Tutoring Research Group (2000) Latent semantic analysis captures causal, goal-oriented and taxonomic structures. In: Proceedings of the annual meeting of the cognitive science society vol **22**, pp 1–6. https://escholarship.org/uc/item/2mw8430f

Guarini M (2006) Particularism and the classification and reclassification of moral cases. IEEE Intell Syst 21(4):22–28

Guarini M (2012) Conative dimensions of machine ethics: a defense of duty. IEEE Trans Affect Comput 3(4):434–442

Hindley M (2009) The voracious pen of Thomas Carlyle. Humanities 30(2):22–26

Hu W (2012) Unsupervised learning of two Bible books: Proverbs and Psalms. Sociol Mind 2(3):325–334. https://doi.org/10.4236/sm.2012.23043

Hu X, Cai Z, Wiemer-Hastings P, Graesser AC, McNamara DS (2011) Strengths, limitations, and extensions of LSA. In: Landauer TK, McNamara DS, Dennis S, Kintsch W Handbook of latent semantic analysis. Routledge, New York, pp 401–425

Iaria A, Schwarz C, Waldinger F (2017) Frontier knowledge and scientific production: evidence from the collapse of international science. Centre for Economic Performance,Discussion Paper No. 1506, http://d.repec.org/n?u=RePEc:ehl:lserod:86599&r=his. Accessed 7 June 2018

Landauer TK (2011) LSA as a theory of meaning. In: Landauer TK, McNamara DS, Dennis S, Kintsch W Handbook of latent semantic analysis. Routledge, New York, pp 3–34

Landauer TK, Foltz PW, Laham D (1998) An Introduction to Latent Semantic Analysis. Discourse Process **25**(2&3):259–284

Landauer TK, McNamara DS, Dennis S, Kintsch W (eds) (2011) Handbook of latent semantic analysis. Routledge, New York

Letsche TA, Berry MW (1997) Large-scale information retrieval with latent semantic indexing. Inf Sci 100(1–4):105–137

Li QV, Ranzato MA, Monga R, Devin M, Chen K, Corrado GS, Dean J, Ng AY (2012) Building high-level features using large scale unsupervised learning. In: Proceedings of the 29th international conference on machine learning. Edinburgh, Scotland, UK

Lin P, Abney K, Bekey GA (eds) (2014) Robot ethics: the ethical and social implications of robotics. The MIT Press, Cambridge

MacIntyre A (1984) After virtue, Second edn. University of Notre Dame Press, Notre Dame

Martin DI, Berry MW (2011) Mathematical foundations behind latent semantic analysis. In: Landauer TK, McNamara DS, Dennis S, Kintsch W Handbook of latent semantic analysis. Routledge, New York, pp 35–55

Mikolov T, Chen K, Corrado G, Dean J (2013a) efficient estimation of word representations in vector space. https://arxiv.org/abs/1301.3781

Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013b) Distributed representations of words and phrases and their compositionality. https://arxiv.org/pdf/1310.4546.pdf

Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130–137

Project Gutenberg (2018). https://www.gutenberg.org/. Accessed 7 June 2018

Salton G, Buckley C (1991) Automatic text structuring and retrieval—experiments in automatic encyclopedia searching. In: Proceedings of the 14th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York

Shiffrin R, Börner K (2004) Mapping knowledge domains. Proc Natl Acad Sci 101(suppl 1):5183–5185

Shirota Y, Chakraborty B (2015) Visual explanation of mathematics in latent semantic analysis. In: IIAI 4th international congress on advanced applied informatics. https://doi.org/10.1109/IIAI-AAI.2015.174

Shlens J (2014) A tutorial on principal component analysis. Google Research. https://arxiv.org/pdf/1404.1100.pdf?utm_content=bufferb37df&utm_medium=social&utm_source=facebook.com&utm_campaign=buffer. Accessed 7 June 2018

Steyvers M, Griffiths T (2011) Probabilistic topic models. In: Landauer TK, McNamara DS, Dennis S, Kintsch W Handbook of latent semantic analysis. Routledge, New York, pp 427–448

Strang G (2016) Introduction to linear algebra, Fifth edn. Wellesley-Cambridge Press, Wellesley

Wallach W, Allen C (2009) Moral Machines: Teaching robots right from wrong. Oxford University Press, Oxford

Wikipedia (2017) Stemming. https://en.wikipedia.org/wiki/Stemming. Accessed 7 June 2018

Wikipedia (2018) Stylometry. https://en.wikipedia.org/wiki/Stylometry. Accessed 7 June 2018

Wolfram Research, Inc. (2017) Mathematica 11