Clustering and Dimensional Reduction for Large Datasets

Jonathan McFadden
University of Washington Tacoma

November 14, 2016

Let $\overline{\overline{\mathbf{L}}} \in \mathbb{Z}^{p \times q}$ be a p by q matrix which describes how often any of $q \in \mathbb{Z}$ n-grams occurs in each of $p \in \mathbb{Z}$ words. That is to say, that the number of times the jth "n-gram" occurs in the ith "word" is represented by $L_{ij} \in \overline{\overline{\mathbf{L}}}$. For simplicity, we will define $\mathfrak{W} \equiv \{w_1, w_2, \cdots, w_p\}$ as the set of all words, with the words represented by the $w_i \in \mathfrak{W}$. Clearly, this implies that $|\mathfrak{W}| = p$. Similarly, we denote the set of all $n-grams$ as $\mathfrak{G} \equiv \{g_1, g_2, \ldots, g_q\}$, with the n-grams represented by the $g_j \in \mathfrak{G}$ and the cardinality of $\mathfrak{G}$ equal to q.

For the words

- politic
- politics
- political
- police
- diner
- dinner
- dinners
- dining
- diners

and the n-grams

- c
- d
- i
- pol
- lit
- tic
- din
- ner
- ing
- tics
- lice
- dine
- ners
- ning
- tical
- inner
- litic
- ining

we have $\overline{\overline{\mathbf{L}}}$ as

$$\overline{\overline{\mathbf{L}}} = \begin{bmatrix} 1 & 0 & 2 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 2 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 2 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{1}$$

For clarity, we have associated the words and n-grams with the rows and columns respectively to give the following table

	c	d	i	pol	lit	tic	din	ner	ing	tics	lice	dine	ners	ning	tical	inner	litic	ining
politic	1	0	2	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0
politics	1	0	2	1	1	1	0	0	0	1	0	0	0	0	0	0	1	0
political	1	0	2	1	1	1	0	0	0	0	0	0	0	0	1	0	1	0
police	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
diner	0	1	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0
dinner	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0
dinners	0	1	1	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0
dining	0	1	2	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1
diners	0	1	1	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0

We now compute the product of $\overline{\overline{\mathbf{L}}}$ with its transpose to obtain

$$\overline{\overline{\mathbf{L}}}_r = \overline{\overline{\mathbf{L}}}\,\overline{\overline{\mathbf{L}}}^{\intercal} = \begin{bmatrix} 9 & 9 & 9 & 4 & 2 & 2 & 2 & 4 & 2 \\ 9 & 10 & 9 & 4 & 2 & 2 & 2 & 4 & 2 \\ 9 & 9 & 10 & 4 & 2 & 2 & 2 & 4 & 2 \\ 4 & 4 & 4 & 4 & 1 & 1 & 1 & 2 & 1 \\ 2 & 2 & 2 & 1 & 5 & 4 & 4 & 4 & 5 \\ 2 & 2 & 2 & 1 & 4 & 5 & 4 & 4 & 4 \\ 2 & 2 & 2 & 1 & 4 & 4 & 5 & 4 & 5 \\ 4 & 4 & 4 & 2 & 4 & 4 & 4 & 9 & 4 \\ 2 & 2 & 2 & 1 & 5 & 4 & 5 & 4 & 6 \end{bmatrix} \tag{2}$$

which we have denoted $\overline{\overline{\mathbf{L}}}_r$. For the sake of clarity and just as above, we will also put this matrix in table form, including labels, to give the following table

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9
w_1 = politic	9	9	9	4	2	2	2	4	2
w_2 = politics	9	10	9	4	2	2	2	4	2
w_3 = political	9	9	10	4	2	2	2	4	2
w_4 = police	4	4	4	4	1	1	1	2	1
w_5 = diner	2	2	2	1	5	4	4	4	5
w_6 = dinner	2	2	2	1	4	5	4	4	4
w_7 = dinners	2	2	2	1	4	4	5	4	5
w_8 = dining	4	4	4	2	4	4	4	9	4
w_9 = diners	2	2	2	1	5	4	5	4	6

This matrix ($\overline{\overline{\mathbf{L}}}_r$) will allow us to group words together based on the strength of their connection. To do this we will construct as set of p tuples where each tuple represents any word linked to the word represented by that tuple. That is to say that for each word, $w_i \in \mathfrak{W}$, we have a set $\mathcal{W}_i$ whose elements are the words to which w_i is linked. Clearly, if w_i is not linked to any other words, then we have $\mathcal{W}_i = \{w_i\}$. Otherwise, we have $\mathcal{W}_i = \{w_i\} \cup \left\{ w_k | w_k \in \mathfrak{W} \text{ and } \overline{\overline{\mathbf{L}}}_{ik} \geq v_t \right\}$, where $v_t \in \mathbb{Z}$ is a threshold value for determining if a word w_k should be included in the set $\mathcal{W}_i$. In order to be more compact, we may rewrite this formal definition of the $\mathcal{W}_i$ as

$$\mathcal{W}_i = \left\{ w_k | w_k = w_i \text{ or } w_k \in \mathfrak{W} \ni \overline{\overline{\mathbf{L}}}_{ik} \geq v_t \right\} \tag{3}$$

We will name connections of the type described in (3) *direct connections*. Additionally, we may use the name *0th order connections* equivalently with *direct connections*.

If we now choose a threshold value of 5 and apply this to our example to obtain the following sets

- $\mathcal{W}_1 = \{w_1, w_2, w_3\}$
- $\mathcal{W}_2 = \{w_2, w_1, w_3\}$
- $\mathcal{W}_2 = \{w_3, w_1, w_2\}$
- $\mathcal{W}_4 = \{w_4\}$
- $\mathcal{W}_5 = \{w_5, w_9\}$
- $\mathcal{W}_6 = \{w_6\}$
- $\mathcal{W}_7 = \{w_7, w_9\}$
- $\mathcal{W}_8 = \{w_8\}$
- $\mathcal{W}_9 = \{w_9, w_5, w_7\}$

From these sets, it is clear that w_1, w_2, & w_3 are all strongly connected to each other; that w_4, w_6, & w_8 are not connected to anything; that w_5 & w_7 strongly connect to w_9; and that w_5 & w_7 do not strongly connect to each other. Recalling that we have defined

- w_1 = politic
- w_2 = politics
- w_3 = political
- w_4 = police
- w_5 = diner
- w_6 = dinner
- w_7 = dinners
- w_8 = dining
- w_9 = diners

It is clear, from $\mathcal{W}_1, \mathcal{W}_2$, and $\mathcal{W}_3$ that

- politic $\longrightarrow \mathcal{W}_1 =$ {politic, politics, political}
- politics $\longrightarrow \mathcal{W}_2 =$ {politics, politics, political}
- political $\longrightarrow \mathcal{W}_3 =$ {political, politic, politics},

from $\mathcal{W}_5, \mathcal{W}_7$, and $\mathcal{W}_9$ that

- diner $\longrightarrow \mathcal{W}_5 =$ {diner, diners}
- dinners $\longrightarrow \mathcal{W}_7 =$ {dinners, diners}
- diners $\longrightarrow \mathcal{W}_9 =$ {diners, diner, dinners},

and from $\mathcal{W}_4, \mathcal{W}_6$, and $\mathcal{W}_8$ that *police*, *dinner*, and *dinning* are not connected to anything.

Finally, we may wish to connect words through shared connections with other words. That is to say, for any words w_l, w_n, and w_m, that if w_n and w_m are both connected to w_l then w_n and w_m may be connected to each other by virtue of their connection to w_l. Thus, we will define another set of p tuples for each word $w_i \in \mathfrak{W}$ such that the tuple represents any word w_m that is connected to w_i through either a direct connection ($w_m \in \mathcal{W}_i$) or through a shared connection with another word w_l, where w_l is directly connected to w_i ($w_l \in \mathcal{W}_i$ and $w_m \notin \mathcal{W}_i$). We will denote these tuples by $\mathcal{W}_i^{\star(1)}$ (*the* (1) *following the* $\star$ *in the superscript will be explained below*) and formally define them by

$$\mathcal{W}_i^{\star(1)} = \mathcal{W}_i \cup \{w_m | w_m \in \mathcal{W}_l \text{ with } w_l \in \mathcal{W}_i\} \tag{4}$$

We will name connections of the type described in (4) *1st order indirect connections*, as opposed to the *direct connections* we defined in (3). We will also use the name *1st order connections* equivalently with *1st order indirect connections*.

We refer to the connections described in (4) as *1st order indirect connections* because only one shared element is required to create the connection. This is why we have the (1) following the $\star$ in the superscript. The (1) denotes that $\mathcal{W}_i^{\star(1)}$ is a set containing only 1st and 0th order connections. It follows that we may create connections based on several shared elements. Similar to our construction in (4), we represent these connections as tuples with

$$\mathcal{W}_i^{\star(r)} = \mathcal{W}_i^{\star(r-1)} \cup \left\{ w_m | w_m \in \mathcal{W}_l^{\star(r-1)} \text{ with } w_l \in \mathcal{W}_i^{\star(r-1)} \right\} \tag{5}$$

as their formal definition for *rTH order connections*. For example, *2nd order connections*, the definition

$$\mathcal{W}_i^{\star(2)} = \mathcal{W}_i^{\star(1)} \cup \left\{ w_m | w_m \in \mathcal{W}_l^{\star(1)} \text{ with } w_l \in \mathcal{W}_i^{\star(1)} \right\}$$

is implied by the expression stated above in (5).

To see this more concretely, let us consider the words w_k, w_l, w_m, and w_n where we have the following connections between them

- w_k is directly connected to w_l
- w_l is directly connected to w_m
- w_m is directly connected to w_n

This implies that the sets

- $\mathcal{W}_k = \{w_k, w_l\}$
- $\mathcal{W}_l = \{w_l, w_k, w_m\}$
- $\mathcal{W}_m = \{w_m, w_l, w_n\}$
- $\mathcal{W}_n = \{w_n, w_m\} = \mathcal{W}_m$

represent the 0th order connections for w_k, w_l, w_m, and w_n. From these 0th order connection sets, we can obtain the following sets

- $\mathcal{W}_k^{\star(1)} = \{w_k, w_l\} \cup \{w_m\}$
- $\mathcal{W}_l^{\star(1)} = \{w_l, w_k, w_m\} \cup \{w_n\}$
- $\mathcal{W}_m^{\star(1)} = \{w_m, w_l, w_n\} \cup \{w_k\}$
- $\mathcal{W}_n^{\star(1)} = \{w_n, w_m\} \cup \{w_l\}$

to represent the 1st order connections for each word. While it is clear that w_l and w_m cannot be connected further, it is also clear that both w_k and w_n may be connected further through 2nd order connections. These second order connections for w_k and w_n can be represented by the sets

- $\mathcal{W}_k^{\star(2)} = \{w_k, w_l, w_m\} \cup \{w_n\}$
- $\mathcal{W}_n^{\star(2)} = \{w_n, w_m, w_l\} \cup \{w_k\}$

While the above sets make clear that no further connections between w_k, w_l, w_m, and w_n are possible, in general this is not necessarily the case. However, since $\mathfrak{W}$ is finite, $\exists r \in \mathbb{Z} \ni \mathcal{W}_i^{\star(r)} = \mathcal{W}_i^{\star(r-1)}, \forall w_i \in \mathfrak{W}$. Further, since $|\mathfrak{W}| = p$, we may bound this r according to $r < p$.

Returning to our persistent example, we see from $\mathcal{W}_1 = \mathcal{W}_2 = \mathcal{W}_2 = \{w_3, w_1, w_2\}$, $\mathcal{W}_4 = \{w_4\}$, $\mathcal{W}_6 = \{w_6\}$, $\mathcal{W}_8 = \{w_8\}$, and $\mathcal{W}_9 = \{w_9, w_5, w_7\}$ that these words may not be further connected. Conversely, from $\mathcal{W}_5 = \{w_5, w_9\}$ and $\mathcal{W}_7 = \{w_7, w_9\}$, it is clear that w_5 and w_7 can be further connected through their shared connection with w_9. This gives us the sets $\mathcal{W}_5^{\star(1)} = \{w_5, w_9\} \cup \{w_7\}$ and $\mathcal{W}_7^{\star(1)} = \{w_7, w_9\} \cup \{w_5\}$ and makes clear that there are no more connections between words, indirect or otherwise.