Webpage Ranking

Eigenvalues and eigenvectors of matrices are extremely important tools with applications in every field of engineering. In this workshop we will study the Google page rank algorithm which revolutionised the search engine, leading to Google becoming one of the largest companies in the world and almost monopolising the internet search market.

You are given the task of analysing internet traffic between Google, Intel, and Microsoft's webpages. After some data analysis, you find that for people visiting Google's page, after each hour, 20% of them would stay on Google's page, 40% would follow a link and go to Intel's page, and 40% to Microsoft's page. For those visiting Intel's page, 50% would stay, 10% would go to Google's page, and 40% to Microsoft's page. For those visiting Microsoft's page, 40% would stay, 30% would go to Google's page, and 30% to Intel's page.

1. Suppose initially there are 1000 visitors to each company's webpage. Assuming there are no additional visitors, how many people are on each company's webpage after 1 hour? **[1 mark]**

   Solution: After one hour, the number of visitors to Google, Intel and Microsoft will be

$$\begin{aligned} G &= 0.2 \times 1000 + 0.1 \times 1000 + 0.3 \times 1000 = 600 \\ I &= 0.4 \times 1000 + 0.5 \times 1000 + 0.3 \times 1000 = 1200 \\ M &= 0.4 \times 1000 + 0.4 \times 1000 + 0.4 \times 1000 = 1200 \end{aligned}$$

2. Let G_n, I_n, M_n denote the number of people on Google, Intel and Microsoft's webpage after n hours respectively, and $\mathbf{x}_n = \begin{bmatrix} G_n \\ I_n \\ M_n \end{bmatrix}$. Write down a matrix $\mathbf{A}$ such that $\mathbf{x}_{n+1} = \mathbf{A}\mathbf{x}_n$. This is called the transition matrix. **[2 marks]**

   Solution: We have

$$\begin{aligned} G_{n+1} &= 0.2G_n + 0.1I_n + 0.3M_n \\ I_{n+1} &= 0.4G_n + 0.5I_n + 0.3M_n \\ M_{n+1} &= 0.4G_n + 0.4I_n + 0.4M_n \end{aligned}$$

   Therefore the matrix equation is

$$\begin{bmatrix} G_{n+1} \\ I_{n+1} \\ M_{n+1} \end{bmatrix} = \begin{bmatrix} 0.2 & 0.1 & 0.3 \\ 0.4 & 0.5 & 0.3 \\ 0.4 & 0.4 & 0.4 \end{bmatrix} \begin{bmatrix} G_n \\ I_n \\ M_n \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} 0.2 & 0.1 & 0.3 \\ 0.4 & 0.5 & 0.3 \\ 0.4 & 0.4 & 0.4 \end{bmatrix}$$

3. Using the transition matrix $\mathbf{A}$ and the help of Matlab or CAS, find out how many people are on each company's webpage after 3 hours. What about after 24 hours? Round your answer to the nearest integer. **[1 marks]**

   Solution: Using Matlab, we find that after 3 hours, the number of visitors will be

$$\begin{bmatrix} G_3 \\ I_3 \\ M_3 \end{bmatrix} = \mathbf{A} \cdot \mathbf{A} \cdot \mathbf{A} \cdot \begin{bmatrix} 1000 \\ 1000 \\ 1000 \end{bmatrix} = \begin{bmatrix} 600 \\ 1200 \\ 1200 \end{bmatrix}$$

After 24 hours, the number of visitors will be

$$\begin{bmatrix} G_{24} \\ I_{24} \\ M_{24} \end{bmatrix} = \mathbf{A}^{24} \cdot \begin{bmatrix} 1000 \\ 1000 \\ 1000 \end{bmatrix} = \begin{bmatrix} 600 \\ 1200 \\ 1200 \end{bmatrix}$$

4. Now suppose initially that Google's webpage receives 2000 visitors, and that each of Intel and Microsoft's webpages receives 500 visitors. How many people are on each company's webpage after 3 and 24 hours respectively? Round your answer to the nearest integer. **[1 marks]**

Solution: Again we can use Matlab to calculate that after 3 hours, the number of visitors will be

$$\begin{bmatrix} G_3 \\ I_3 \\ M_3 \end{bmatrix} = \mathbf{A} \cdot \mathbf{A} \cdot \mathbf{A} \cdot \begin{bmatrix} 2000 \\ 500 \\ 500 \end{bmatrix} = \begin{bmatrix} 600 \\ 1200 \\ 1200 \end{bmatrix}$$

After 24 hours, the number of visitors will be

$$\begin{bmatrix} G_{24} \\ I_{24} \\ M_{24} \end{bmatrix} = \mathbf{A}^{24} \cdot \begin{bmatrix} 2000 \\ 500 \\ 500 \end{bmatrix} = \begin{bmatrix} 600 \\ 1200 \\ 1200 \end{bmatrix}$$

You notice that the numbers of visitors to the company's webpages seem to stabilise as time passes, regardless of where the 3000 visitors start initially! Let's investigate this further.

5. Show that if the number of visitors stabilises to some time independent constant $\mathbf{x} = \begin{bmatrix} G \\ I \\ M \end{bmatrix}$, then $\mathbf{x}$ is in fact an eigenvector of $\mathbf{A}$ with eigenvalue 1. **[2 marks]**

Solution: Since $\mathbf{x}$ is stable, it should be unchanged under multiplication by the transition matrix, i.e., $\mathbf{Ax} = \mathbf{x}$. But this is the definition of $\mathbf{x}$ being an eigenvector of $\mathbf{A}$ with eigenvalue 1.

6. Write down the characteristic equation and find the eigenvalues of the transition matrix $\mathbf{A}$. **[2 marks]**

Solution: The characteristic equation is

$$\begin{aligned} \det(\mathbf{A} - \lambda \mathbf{I}) &= \det \begin{bmatrix} 0.2-\lambda & 0.1 & 0.3 \\ 0.4 & 0.5-\lambda & 0.3 \\ 0.4 & 0.4 & 0.4-\lambda \end{bmatrix} \\ &= -\lambda^3 + 1.1\lambda^2 - 0.1\lambda \end{aligned}$$

This factorises into

$$-\lambda^3 + 1.1\lambda^2 - 0.1\lambda = -\lambda(\lambda - 1)(\lambda - 0.1)$$

Therefore the eigenvalues are 1, 0.1, and 0.

7. For each eigenvalue of $\mathbf{A}$, find the corresponding eigenvector(s). **[4 marks]**

Solution: For $\lambda = 1$,

$$\mathbf{A} - \lambda \mathbf{I} = \begin{bmatrix} -0.8 & 0.1 & 0.3 \\ 0.4 & -0.5 & 0.3 \\ 0.4 & 0.4 & -0.6 \end{bmatrix}$$

Swap R_1 and R_2,

$$\begin{bmatrix} 0.4 & -0.5 & 0.3 \\ -0.8 & 0.1 & 0.3 \\ 0.4 & 0.4 & -0.6 \end{bmatrix}$$

$R_2 = R_2 + 2R_1$, $R_3 = R_3 - R_1$,

$$\begin{bmatrix} 0.4 & -0.5 & 0.3 \\ 0 & -0.9 & 0.9 \\ 0 & 0.9 & -0.9 \end{bmatrix}$$

$R_3 = R_3 + R_2$,

$$\begin{bmatrix} 0.4 & -0.5 & 0.3 \\ 0 & -0.9 & 0.9 \\ 0 & 0 & 0 \end{bmatrix}$$

M is the free variable. Setting $M = t$, we find $-0.9I + 0.9t = 0$, so $I = t$. And $0.4G - 0.5t + 0.3t = 0$, so $G = \frac{1}{2}t$. Therefore the solutions space is $\begin{bmatrix} \frac{1}{2}t \\ t \\ t \end{bmatrix} = t \begin{bmatrix} \frac{1}{2} \\ 1 \\ 1 \end{bmatrix}$, and $\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$ is an eigenvector.

For $\lambda = 0.1$,

$$\mathbf{A} - \lambda \mathbf{I} = \begin{bmatrix} 0.1 & 0.1 & 0.3 \\ 0.4 & 0.4 & 0.3 \\ 0.4 & 0.4 & 0.3 \end{bmatrix}$$

$R_3 = R_3 - R_2$,

$$\begin{bmatrix} 0.1 & 0.1 & 0.3 \\ 0.4 & 0.4 & 0.3 \\ 0 & 0 & 0 \end{bmatrix}$$

$R_2 = R_2 - 4R_1$,

$$\begin{bmatrix} 0.1 & 0.1 & 0.3 \\ 0 & 0 & -0.9 \\ 0 & 0 & 0 \end{bmatrix}$$

I is the free variable. Setting $I = t$, we find $-0.9M = 0$, so $M = 0$. And $0.1G + 0.1t = 0$, so $G = -t$. Therefore the solutions space is $\begin{bmatrix} -t \\ t \\ 0 \end{bmatrix} = t \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$, and $\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$ is an eigenvector.

For $\lambda = 0$,

$$\mathbf{A} - \lambda \mathbf{I} = \begin{bmatrix} 0.2 & 0.1 & 0.3 \\ 0.4 & 0.5 & 0.3 \\ 0.4 & 0.4 & 0.4 \end{bmatrix}$$

$R_2 = R_2 - 2R_1, R_3 = R_3 - 2R_1$,

$$\begin{bmatrix} 0.2 & 0.1 & 0.3 \\ 0 & 0.3 & -0.3 \\ 0 & 0.2 & -0.2 \end{bmatrix}$$

$R_3 = R_3 - \frac{2}{3}R_2$,

$$\begin{bmatrix} 0.2 & 0.1 & 0.3 \\ 0 & 0.3 & -0.3 \\ 0 & 0 & -0 \end{bmatrix}$$

M is the free variable. Setting $I = t$, we find $0.3I - 0.3t = 0$, so $I = t$. And $0.2G + 0.1t + 0.3t = 0$, so $G = -2t$. Therefore the solutions space is $\begin{bmatrix} -2t \\ t \\ t \end{bmatrix} = t \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$, and $\begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$ is an eigenvector.

8. Diagonalise the matrix $\mathbf{A}$, i.e., find a diagonal matrix $\mathbf{D}$ and an invertible matrix $\mathbf{V}$ such that $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$. **[2 marks]**

   Solution: To diagonalise $\mathbf{A}$, the diagonal matrix $\mathbf{D}$ is the matrix of eigenvalues and $\mathbf{V}$ is a matrix of corresponding eigenvectors. Hence

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{V} = \begin{bmatrix} 1 & -1 & -2 \\ 2 & 1 & 1 \\ 2 & 0 & 1 \end{bmatrix}$$

   and so

$$\begin{bmatrix} 0.2 & 0.1 & 0.3 \\ 0.4 & 0.5 & 0.3 \\ 0.4 & 0.4 & 0.4 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -2 \\ 2 & 1 & 1 \\ 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & -1 & -2 \\ 2 & 1 & 1 \\ 2 & 0 & 1 \end{bmatrix}^{-1}$$

9. Show that $\mathbf{A}^n = \mathbf{V}\mathbf{D}^n\mathbf{V}^{-1}$ in terms of $\mathbf{D}$ and $\mathbf{V}$. **[1 mark]**

   Solution: Writing out the product we have

$$\begin{aligned} \mathbf{A}^n &= (\mathbf{V}\mathbf{D}\mathbf{V}^{-1})^n = (\mathbf{V}\mathbf{D}\mathbf{V}^{-1})(\mathbf{V}\mathbf{D}\mathbf{V}^{-1})\cdots(\mathbf{V}\mathbf{D}\mathbf{V}^{-1}) \\ &= \mathbf{V}\mathbf{D}(\mathbf{V}^{-1}\mathbf{V})\mathbf{D}(\mathbf{V}^{-1}\mathbf{V})\cdots\mathbf{D}\mathbf{V}^{-1} = \mathbf{V}\mathbf{D}^n\mathbf{V}^{-1} \end{aligned}$$

10. For the particular $\mathbf{D}$ you have, what happens to $\mathbf{D}^n$ as n becomes large? Use your observation and the previous part to explain why the numbers of visitors to the company's webpages stabilise as time passes, regardless of where the 3000 visitors start initially. **[3 marks]**

   Solution: Since $\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ is a diagonal matrix, we have

$$\mathbf{D}^n = \begin{bmatrix} 1^n & 0 & 0 \\ 0 & (0.1)^n & 0 \\ 0 & 0 & 0^n \end{bmatrix}$$

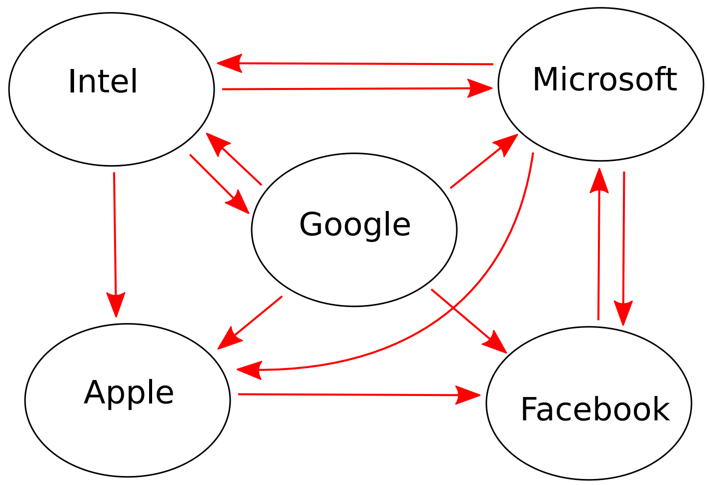
   Since the magnitudes of 0.1 is less than 1, $(0.1)^n$ tends to 0 as n becomes large. Hence $\mathbf{D}$ approaches $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ for large values of n. Therefore for any initial distribution G_0, I_0, M_0 of visitors where $G_0 + I_0 + M_0 = 3000$,

$$\mathbf{x}_n = \mathbf{A}^n \begin{bmatrix} G_n \\ I_n \\ M_n \end{bmatrix} = \mathbf{V}\mathbf{D}^n\mathbf{V}^{-1} \begin{bmatrix} G_n \\ I_n \\ M_n \end{bmatrix}$$

   Since $\mathbf{D}^n$ approaches $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and $\mathbf{V} = \begin{bmatrix} 1 & -1 & -2 \\ 2 & 1 & 1 \\ 2 & 0 & 1 \end{bmatrix}$, $\mathbf{x}_n$ approaches

$$\begin{aligned} &\begin{bmatrix} 1 & -1 & -2 \\ 2 & 1 & 1 \\ 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & -1 & -2 \\ 2 & 1 & 1 \\ 2 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} G_n \\ I_n \\ M_n \end{bmatrix} \\ =&\frac{1}{5} \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{bmatrix} \begin{bmatrix} G_n \\ I_n \\ M_n \end{bmatrix} = \begin{bmatrix} \frac{1}{5}(G_n + I_n + M_n) \\ \frac{2}{5}(G_n + I_n + M_n) \\ \frac{2}{5}(G_n + I_n + M_n) \end{bmatrix} = \begin{bmatrix} 600 \\ 1200 \\ 1200 \end{bmatrix} \end{aligned}$$

You can now analyse a more complicated network of webpages. The diagram below shows the webpages of Apple, Facebook, Google, Intel and Microsoft. A directed arrow indicates that there is a link on the starting webpage to the ending webpage. Suppose that after each hour, a person visiting a webpage has an equal chance of staying on the webpage or following any one of its links. For example, in this diagram, anyone visiting Google has a 20% chance of being on Apple, Facebook, Google, Intel and Microsoft's webpage after an hour.



11. Write down the transition matrix $\mathbf{A}$ for this network of webpages, ordering the companies alphabetically for your variables. **[1 mark]**

    Solution: Reading off the diagram, ordering the companies in alphabetical order, we have

$$\mathbf{A} = \begin{bmatrix} 0.5 & 0 & 0.2 & 0.25 & 0.25 \\ 0.5 & 0.5 & 0.2 & 0 & 0.25 \\ 0 & 0 & 0.2 & 0.25 & 0 \\ 0 & 0 & 0.2 & 0.25 & 0.25 \\ 0 & 0.5 & 0.2 & 0.25 & 0.25 \end{bmatrix}$$

12. For the same reason as in Question (10), the number of visitors to each page will stabilise regardless of initial distribution. With the help of Matlab or CAS, rank the webpages by the stabilised number of visitors in descending order. **[2 marks]**

    Solution: Using Matlab or a calculator, we find an eigenvector of $\mathbf{A}$ with eigenvalue 1 to be

$$\mathbf{v} = \begin{bmatrix} 32 \\ 56 \\ 5 \\ 16 \\ 44 \end{bmatrix}$$

    Therefore the page ranking is Facebook, Microsoft, Apple, Intel, Google.

This is essentially Google's page ranking algorithm. It uses the links between webpages to determine which pages are likely be the most important and of the most interest to users. Of course, the actual algorithm is more sophisticated and remains proprietary, but the basic idea remain the same: it's simply eigenvalues and eigenvectors!

There is also 1 additional mark given for the quality of the English and 1 additional mark for correct mathematical notation. These marks are easy to obtain but the markers will be instructed to be strict in awarding these marks.