MCA - HW3

-Harkishan Singh (2017233)

Question 1:

Steps of algorithm :

1) Imported ‘abc’ corpus. It contains more than 500k words. Removed stop words.
2) Randomly choose 10k words out of ‘abc’ corpus. **One hot encoded** those 10k words.
3) Created trainable tuples with window size of 1 from the corpus.
4) Initialised my neural network with following layers :

    A) Input layer : 10k nodes

    B) Hidden layer : 300 nodes

    C) Output layer : 10k nodes

Binary Cross entropy loss is used.

5) Trained the model.
6) Choose 20 random words from the corpus for testing => one hot encoded them.
7) Predicted the value vector of testing words.
8) Applied PCA (Principal Component Analysis) to reduce the size of testing vectors to a vector of dimension = 2 (for plotting purpose).
9) Plotted the 2 dimensional vector.

Below shown is the list of testing words selected and their PCA 2 dimensions value :

\

	word	x	y
0	access	-2.646354	-2.885437
1	scientists	10.056523	9.584141
2	shells	-14.896098	-6.884211
3	Mechanism	0.098784	-6.258768
4	would	-12.059207	-11.021449
5	exposure	2.176972	16.664717
6	used	1.988377	9.313240
7	gas	-13.825939	12.995003
8	sports	15.357718	2.921603
9	clockwork	4.701605	-13.778342
10	chew	-10.196159	-5.472467
11	called	8.861832	10.793338
12	safety	-5.334979	-2.981449
13	domestic	8.309322	-8.427451
14	mango	10.686925	-8.899676
15	suggest	-2.756380	5.982088
16	long	-0.702356	-8.127085
17	So	12.096659	-1.813099
18	In	3.221068	-2.069785
19	north	-15.138387	10.365196

When plotted, plot looks like :