MCA Assignment 3

By
Rohan Chhokra
2016080

Part 1

The algorithm comes in two flavours - CBOW and Skip Grams which are closely intertwined with each other. I followed Skip Gram.

Skip Gram works on trying to associate the relations between words occurring within a specific distance(called window size) of each other.

Let's say the sentence is A1,A2,A3,A4,A5... A10
Let's say we're interested in A4, for a context window of 2, we have A2 and A3 at its left hand side and A5 and A6 at its right hand side. Our primary task is to make a neural network which learns this association between a 'center' word(in this case A4) and it's context words (in this case A2/A3/A5/A6).

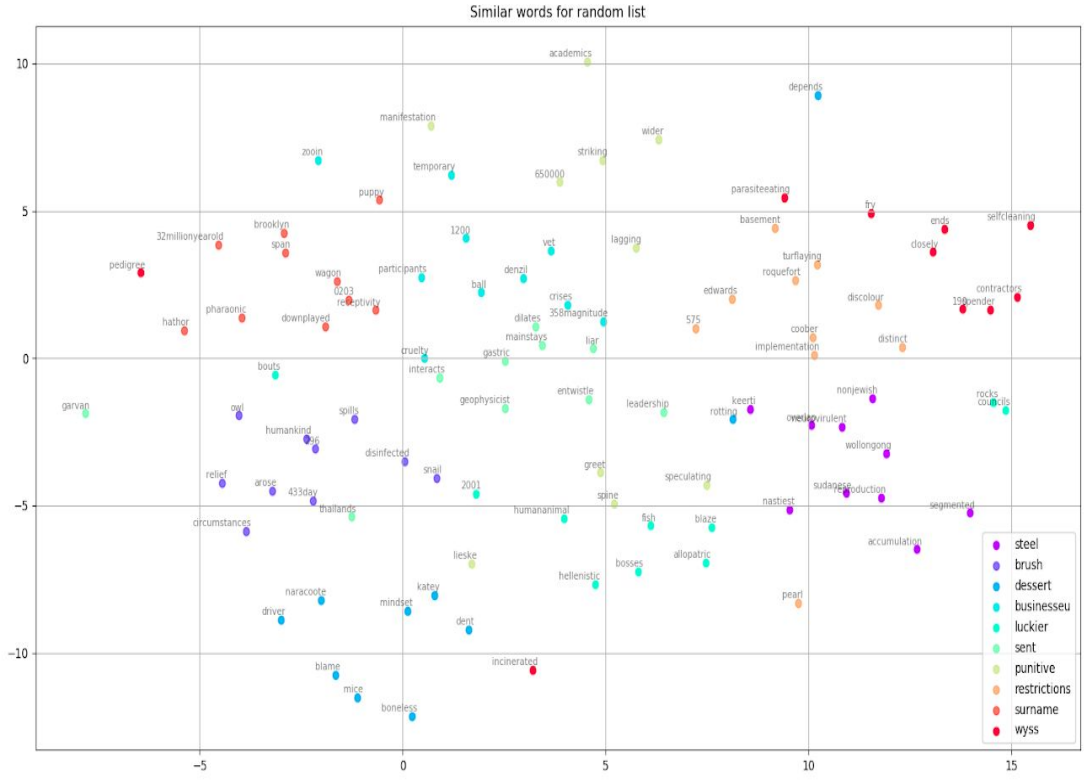
We train our model to predict the context words for a given center word. Now this can happen in a couple of 'sub-flavours' . We can use a softmax layer to predict the

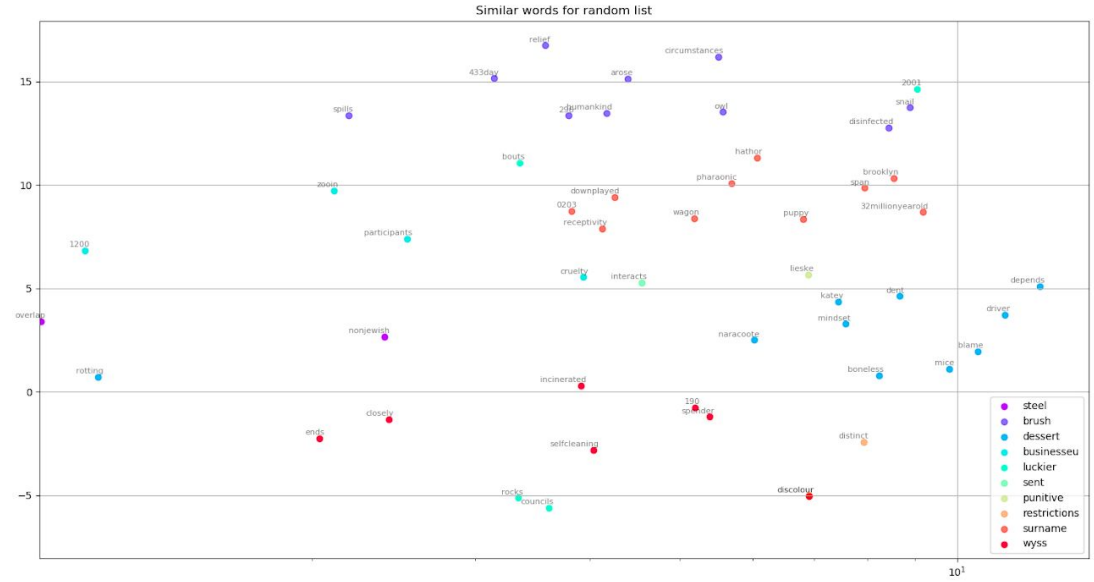
most probably layer or we can use a log-sigmoid layer to predict a ‘yes’ or ‘no’ whether a given context and center word are associated with each other or not. I follow the latter.

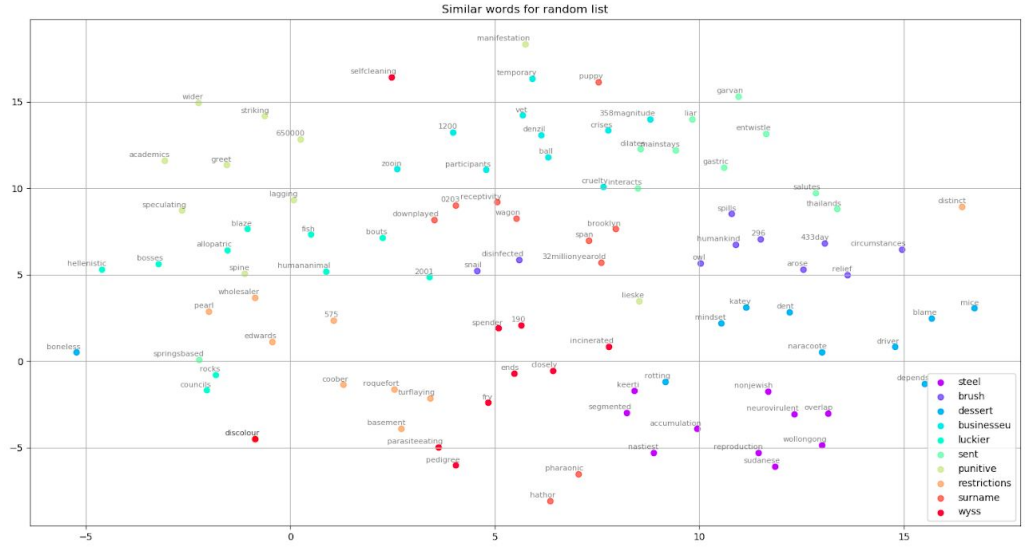
I have a hidden embedding layer of size [vocabulary_size,embedding_size] . My choice of embedding size is 60. For every pair on input - both of them are passed through the embedding layer and then their dot products are calculated using the torch.matmul function. This result is passed through a sigmoid layer to determine whether they are associated or not.

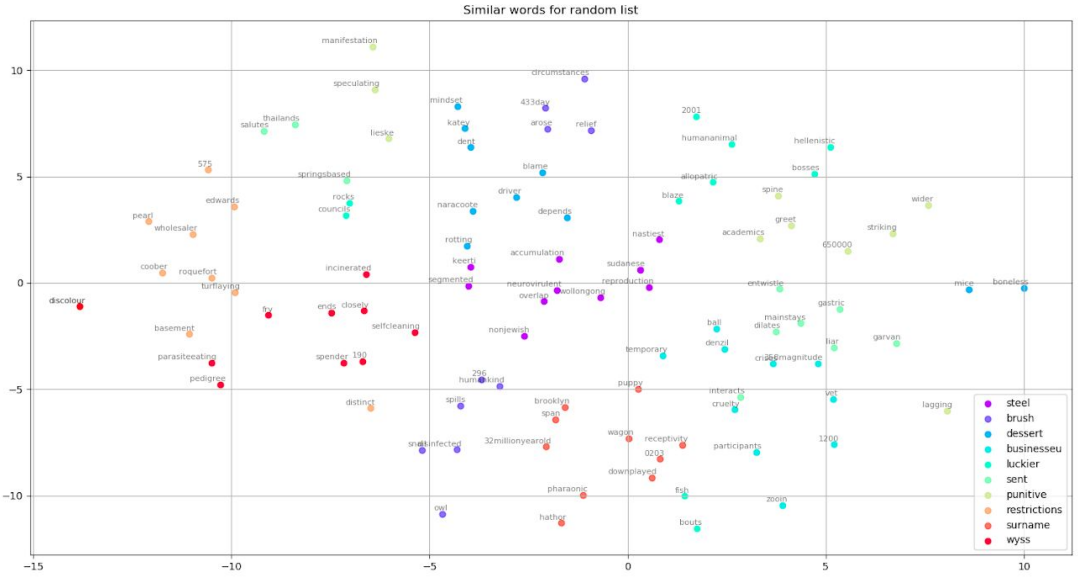
I trained 5 epochs due to computational difficulties and the plots are as follows.

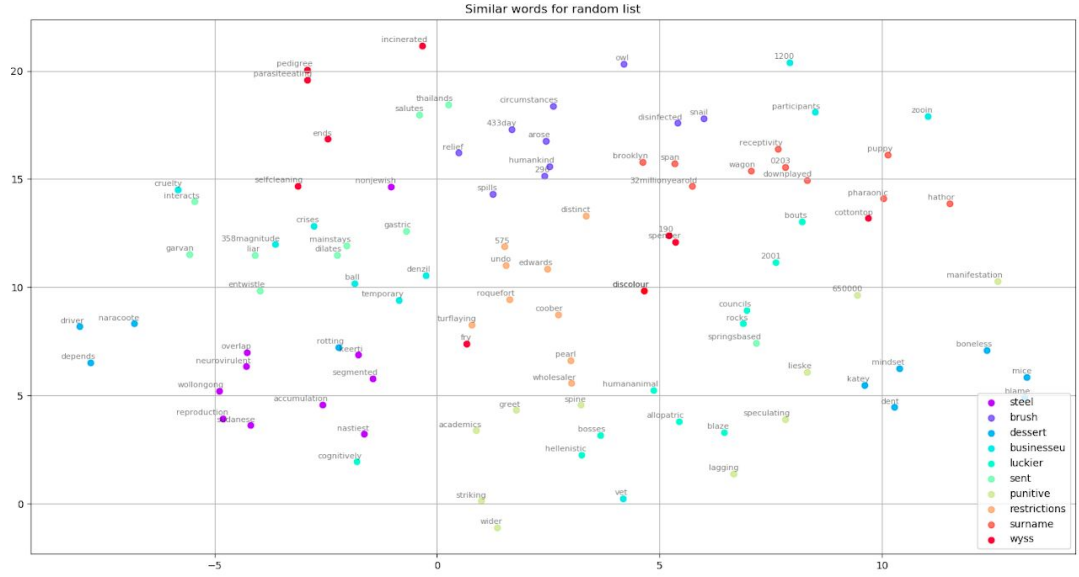
The results are for 10 random words and their closest respective words(using Cosine Similarity logic).


Similar words for random list
academics
depends
manifestation
wider
striking
zoom
temporary
650000
parasiteeating
puppy
fry
basement
ends
selfcleaning
brooklyn
32millionyearold
span
1200
vet
lagging
closely
turflaying
pedigree
wagon
participants
denzil
roquefort
0203
ball
edwards
discolour
contractors
crises
pharaonic
358magnitude
dilates
575
hathor
downplayed
coober
mainstays
liar
implementation
distinct
cruelty
gastric
bouts
interacts
nonjewish
entwistle
geophysicist
leadership
keerti
rotting
garvan
owl
spills
humankind
disinfected
wollongong
snail
greet
speculating
relief
arose
2001
sudanese
spine
nastiest
segmented
433day
thailands
humananimal
fish
blaze
circumstances
accumulation
lieske
allopatric
bosses
hellenistic
naracoote
katey
pearl
mindset
driver
dent
incinerated
blame
mice
boneless
steel
brush
dessert
businesseu
luckier
sent
punitive
restrictions
surname
wyss


Similar words for random list
relief
circumstances
433day
arose
2001
snail
spills
humanking
owl
disinfected
bouts
hathor
brooklyn
zoom
pharaonic
span
downplayed
32millionyearold
wagon
puppy
receptivity
participants
1200
cruelty
interacts
lieske
depends
katey
dent
driver
overlay
mindset
nonjewish
naracoorte
blame
rotting
boneless
mice
incinerated
closely
ends
selfcleaning
distinct
rocks
councils
discolour
steel
brush
dessert
businesseu
luckier
sent
punitive
restrictions
surname
wyss


Similar words for random list
manifestation
selfcleaning
temporary
puppy
garvan
wider
striking
vet
358magnitude
liar
1200
crises
entwistle
650000
denzil
dilate
mainstays
academics
greet
ball
gastric
zoom
participants
cruelty
interacts
salutes
lagging
receptivity
thailands
distinct
speculating
downplayed
wagon
spills
blaze
fish
bouts
brooklyn
span
humanking
296
433day
circumstances
allopatric
disinfected
32millionyearold
owl
arose
relief
bosses
hellenistic
spine
humananimal
snail
2001
wholesaler
lieske
pearl
katey
dent
mice
575
spender
190
blame
mindset
edwards
incinerated
naracoorte
driver
boneless
springsbased
rocks
ends
closely
rotting
depends
councils
coober
roquefort
keerti
nonjewish
burflaying
segmented
neurovirulent
overlay
basement
accumulation
discolour
parasiteeating
nastiest
reproduction
wollongong
pedigree
sudanese
pharaonic
hathor
steel
brush
dessert
businesseu
luckier
sent
punitive
restrictions
surname
wyss


Similar words for random list
steel
brush
dessert
businesseu
luckier
sent
punitive
restrictions
surname
wyss


Similar words for random list
steel
brush
dessert
businesseu
luckier
sent
punitive
restrictions
surname
wyss

Due to poor visibility here images are being added to the repository. They are in order of epochs.

I have referred to the following resources -
https://www.kdnuggets.com/2018/04/implementing-deep-learning-methods-feature-engineering-text-data-skip-gram.html
http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
https://github.com/jojonki/word2vec-pytorch
https://towardsdatascience.com/word2vec-skip-gram-model-part-1-intuition-78614e4d6e0b
https://medium.com/datadriveninvestor/word2vec-skip-gram-model-explained-383fa6ddc4ae

Part 2

I got the following results in Part 2
Retrieval with Relevance Feedback
MAP: 0.6349721389266573

Retrieval with Relevance Feedback and query expansion
MAP: 0.5996370763003258

In the code itself I used multiple iterations - 1 and 3 and got a result of 0.61 and 0.634 . The increase was expected along with more iterations as we see.

In the second part of the second question query expansion surprisingly did not increase. A reason for this might be that when words similar to the most important words of each query messes up the inferred meaning of the query and hence results decrease.

Resources referred -

https://nlp.stanford.edu/IR-book/pdf/09expand.pdf