MCA Assignment-3
Arushi Chauhan
2016019

Q1: Terminology:

1. Context and center word - it refers to the window consisting of n words to the right and left of a chosen center word. For instance, in the sentence “a lion crossed the road”, if “crossed” is a center word and window size is 2, then context would contain the words [“a”,”lion”,”the”,”road”].
2. Bag of words model - It takes a ‘multiset’ of words in a sentence, discarding the positional information and grammar.

Library used: Pytorch

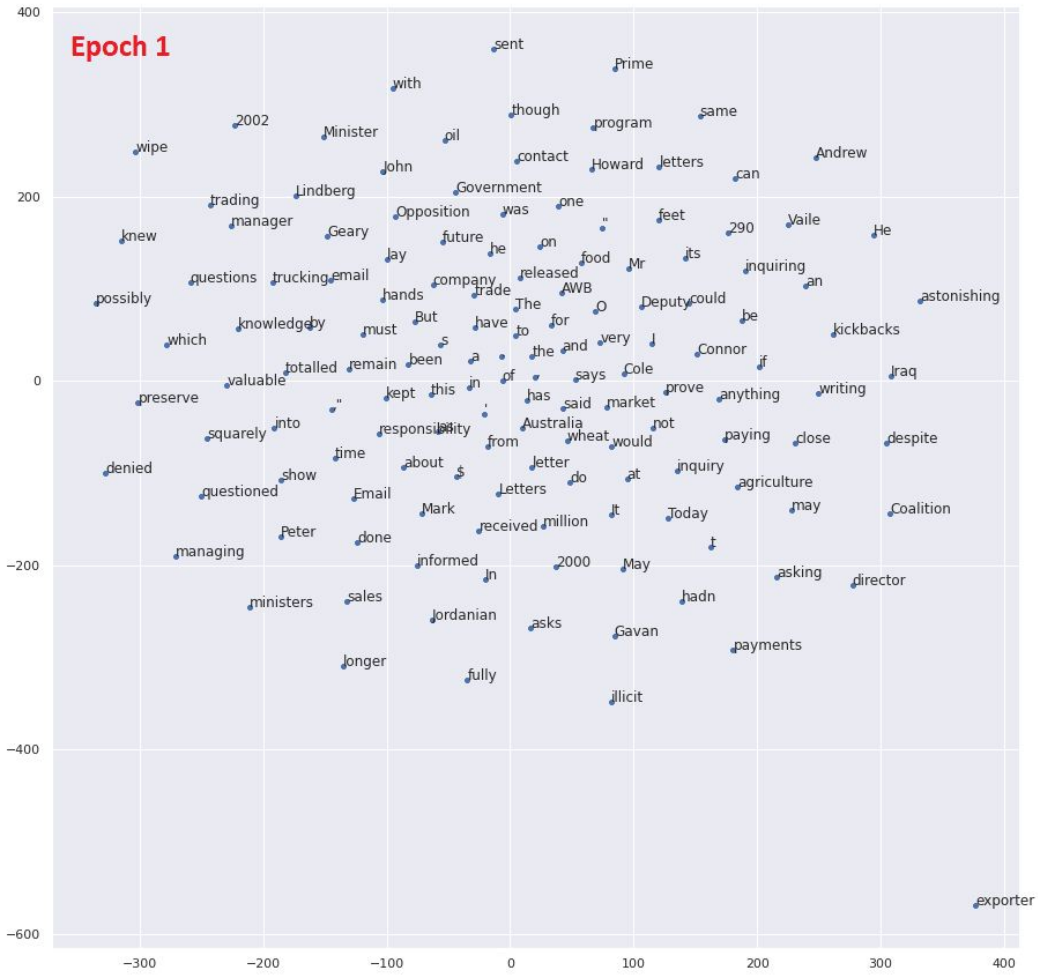
Word2Vec model:

The Word2Vec model consists of two components - CBOW (continuous bag of words) and skipgram. The CBOW part takes in context as input and predicts the center word. The order of the words in the context does not matter, that is why it is a bag of words model. The skipgram takes the center word as input and outputs the most relevant context of the words. Both of these are shallow models, i.e. they generally have one hidden layer to compute the vector representations.

Implementation details:

For this assignment, I have used one embedding layer and one hidden layer in both CBOW and SkipGram models. The models were trained for 50 epochs each. In case of CBOW model, the context words were given as inputs. These context words were passed through embedding layer and the mean of these embedded vectors was taken to get a mean representation of the context. This mean vector was then passed through an FC layer to get a vector of size the length of vocabulary. Taking log softmax of this vector gives the word which the CBOW model predicts as the center word as in one-hot encoding. For skipgram model, during training the center word was passed through the embedding layer and the rest of the training proceeded similar as above.

The TSNE plots for epochs 10, 30 and 50 are as shown at the end of report in figures 1,2, and 3. We can also see the compiled TSNE plots for each epoch in the following GIF. From these plots, we can notice that words that occur frequently together in the corpus move closer as epochs increase. For instance, in epoch 50 (figure 3), we can see [“Howard”,”Program”], [“”minister”,”denied”, “email”], [“possibly”,”manager”,”trading”] are occurring together.


Epoch 1
sent
Prime
with
though
same
2002
Minister
oil
program
wipe
contact
Howard
letters
Andrew
John
can
Government
Lindberg
trading
was
one
Opposition
feet
290
Vaile
He
manager
knew
Geary
future
he
on
lay
food
Mr
its
inquiring
questions
trucking
email
company
released
an
possibly
hands
trade
AWB
The
O
Deputy
could
astonishing
knowledge
by
But
have
for
be
must
to
kickbacks
which
s
and
very
I
Connor
totalled
remain
been
a
the
if
Iraq
valuable
of
says
Cole
0
preserve
kept
this
in
has
said
market
prove
anything
writing
"
'
squarely
into
responsibility
Australia
not
paying
close
despite
from
wheat
would
time
about
letter
inquiry
denied
show
$
do
at
agriculture
questioned
Email
Letters
Mark
It
Today
may
Coalition
Peter
done
received
million
t
managing
informed
2000
May
asking
director
In
ministers
sales
hadn
Jordanian
asks
Gavan
payments
longer
fully
illicit
−200
−400
exporter
−600
−300
−200
−100
0
100
200
300
400

Q2: The following values were reported after relevance feedback and query expansion were implemented:

MAP values	**Relevance Feedback**	**Relevance Feedback with Query Expansion**
1 iteration	0.6544	0.6864
3 iterations	0.9301	0.9720

Steps followed in relevance feedback:

1. Load the ground truth from the file MED.REL.
2. For each query, follow these steps:
    a. Using similarity matrix passed as argument, compute the top 10 documents that are perceived as relevant (according to similarity computed of queries and documents)
    b. Classify the perceived relevant documents as relevant and non-relevant using ground truth as comparison.
    c. Sum all the relevant documents and non-relevant documents separately and multiply the sum with alpha and beta respectively.
    d. Divide the above vectors with length of relevant and non-relevant documents respectively.
    e. Update the query by adding the modified sum of relevant documents and subtracting the modified sum of non-relevant documents from it.

Steps followed in relevance feedback with query expansion:

1. Load the ground truth from the file MED.REL.
2. For each query, follow these steps:
    a. From the ground truth, collect top n terms (those having highest TF-IDF values)
    b. Add the TF-IDF values of these top n terms to the query.

Role of alpha and beta:

Alpha and beta control the sum of relevant and non-relevant documents that are added to the query. The following values were obtained for different values of alpha and beta:

Alpha	Beta	MAP Value
0.75	0.15	0.65
0.75	0	0.62
0.5	0.15	0.63
0	0.15	0.51
0.9	0.8	0.66

We can see that the influence of alpha is greater than the influence of beta on the MAP value, demonstrated by the huge gap in MAP values when alpha = 0 versus when beta = 0. Furthermore, the values tend to remain mostly near 0.64 for most values of alpha and beta pairs.

Q1 figures:

Figure 1

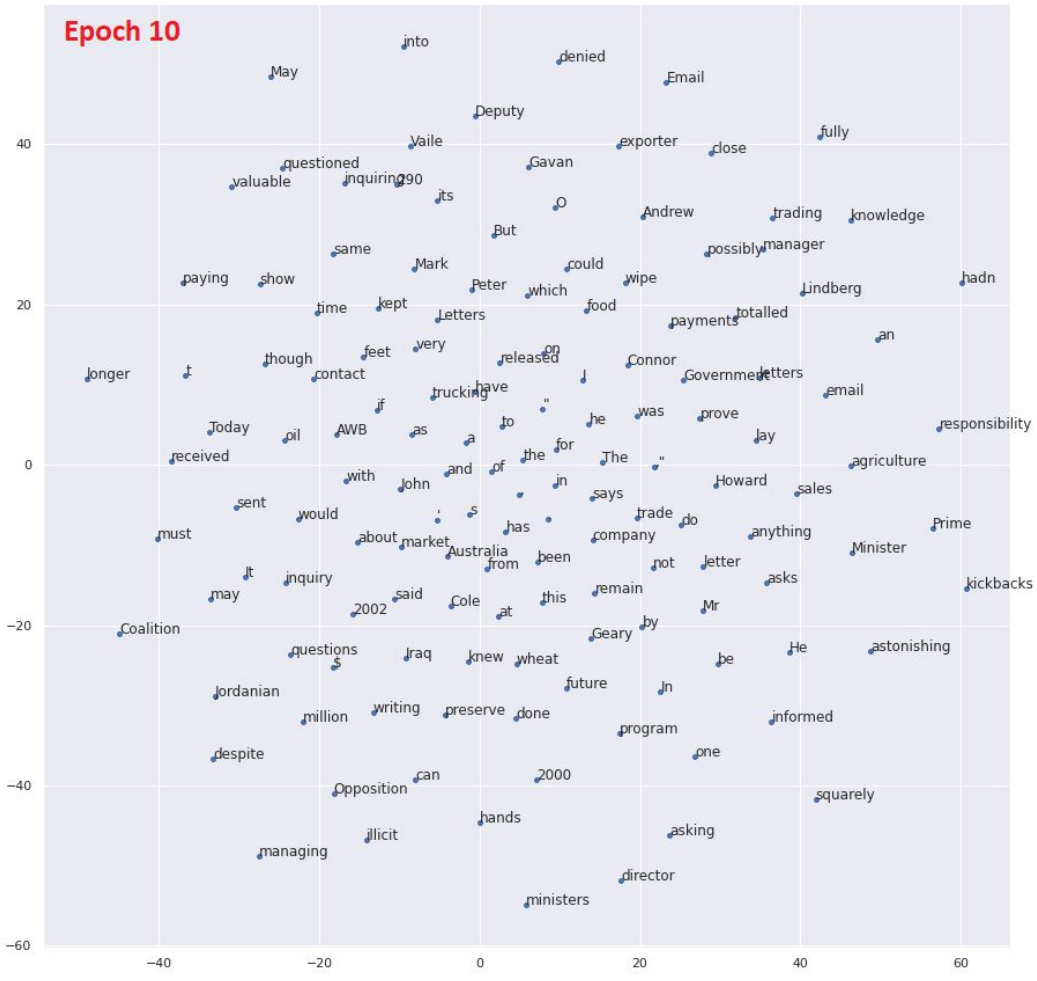


Figure 2

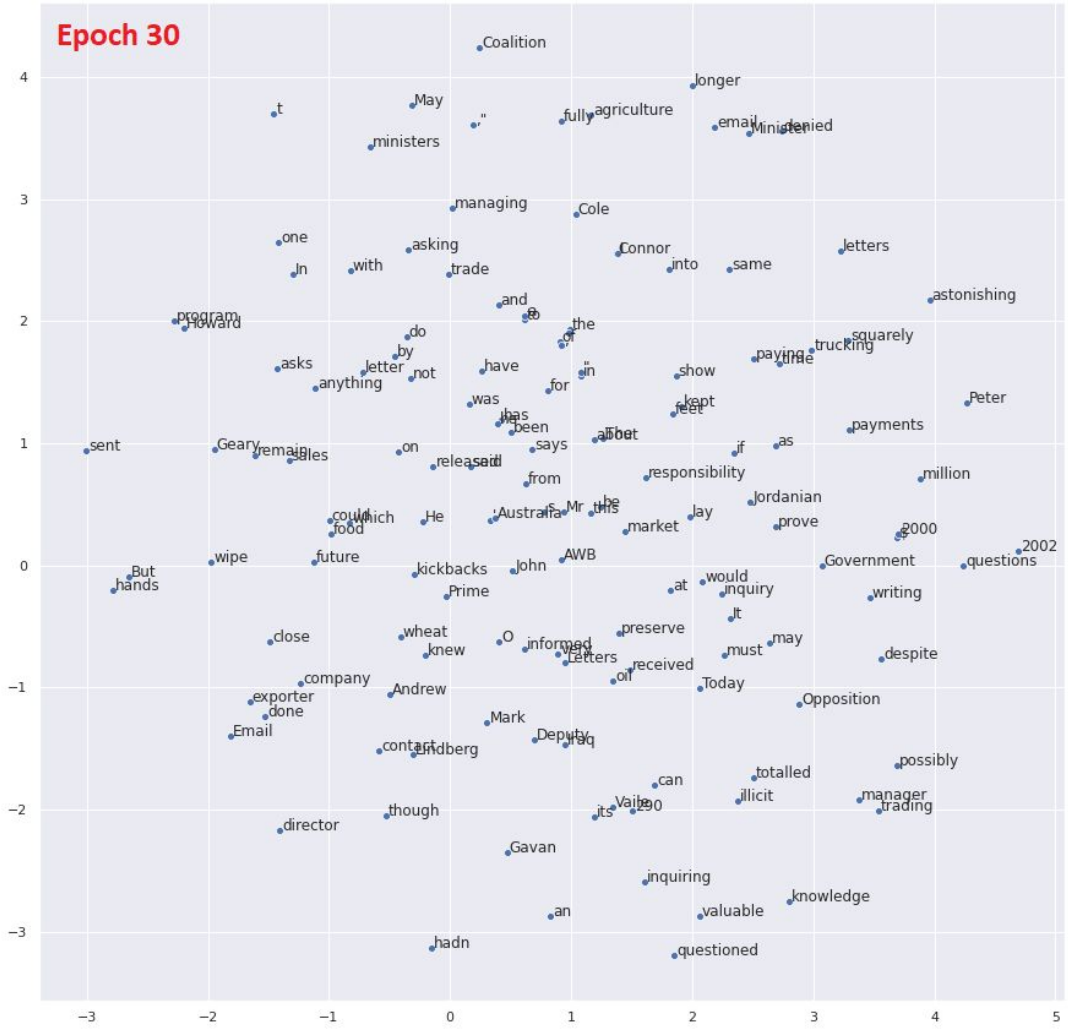

Epoch 30
Coalition
longer
May
t
fully
agriculture
email
Minister
denied
ministers
managing
Cole
one
asking
Connor
letters
In
with
trade
into
same
astonishing
and
program
Howard
do
the
squarely
by
trucking
paying
asks
letter
have
in
show
anything
not
for
kept
Peter
was
has
been
payments
sent
Geary
remain
sales
on
says
if
as
released
said
responsibility
million
from
Jordanian
could
which
He
Australia
Mr
this
Jay
market
prove
2000
food
2002
wipe
future
AWB
Government
questions
But
kickbacks
John
would
hands
Prime
at
inquiry
writing
It
close
wheat
O
preserve
may
knew
informed
must
despite
letters
received
company
oil
Today
Andrew
exporter
Opposition
done
Mark
Email
Deputy
contact
Bridberg
possibly
totalled
can
illicit
manager
trading
director
though
its
290
Gavan
inquiring
knowledge
an
valuable
hadn
questioned

Figure 3


Epoch 50
director
sent
Coalition
asking
with
hands
But
May
,"
ministers
managing
one
In
Geary
remain
trade
and
a
to
program
Howard
sales
Email
exporter
done
by
do
of
the
close
which
future
on
have
in
into
company
anything
food
about
same
Jonger
asks
released
said
has
for
he
"
I
wheat
says
was
Connor
email
denied
contact
Andrew
this
not
Minister
though
letters
knew
He
from
Australia
s
letter
show
trucking
The
responsibility
squarely
time
preserve
at
Cole
Mr
be
John
Prime
market
AWB
would
O
astonishing
kickbacks
Mark
inquiry
very
payments
received
It
as
if
lay
must
oil
kept
prove
Government
t
Today
paying
Letters
informed
may
Jordanian
Peter
2002
Iraq
feet
writing
$
Deputy
despite
fully
agriculture
questions
Lindberg
Opposition
2000
million
can
totalled
Vaile
290
its
wipe
illicit
manager
possibly
trading
inquiring
an
hadn
knowledge
valuable
Gavan
questioned