Introduction to NLP (CS7.401)

Spring 2022, IIIT Hyderabad
Assignment 2

Abhinav S Menon

1 Theory

Negative sampling consists of augmenting the training data with examples *not* present in the corpus. It is done in order that the model is not trained on data that contains only positive results (and consequently return only positive results on any input).

For example, if we are training a CBOW model without negative sampling, we compare the average of the context words to the focus word, and the difference becomes our error.
In contrast, when we add NS, we treat the cosine similarity as the model's confidence that the given focus word occurs in the context of the given window. Thus the training output would be 1 for all samples that we retrieve from the corpus. We then add (context, focus) pairs that are absent from the corpus, in order to provide the model with 0-samples in its training data.

2 Analysis

The top 10 closest words for *camera* in the SVD-based model are:

```
clickshut
soo
soaked
that.tl
second.cons
9/2010-
excursion
finished
size.camera
below============================it
```

We can see that these have very little meaning, except the first one. Some of them are the result of poor tokenization as well; the 9th one is almost a repetition

for this reason.

The pre-trained word vectors return the following top 10 closest words:

```
cameras
Wagging finger
camera lense
camcorder
Camera
Canon digital SLR
Cameras
tripod
EyeToy USB
videocamera
```

(obtained from **here**)
These results prove that the embeddings are similar, as they should be, but this particular query is of very little use as text normalisation has not been carried out (so only variants of the same word reappear, rather than semantically close tokens).