CS5560 Knowledge Discovery and Management
Problem Set 4
June 26 (T), 2017

Name: SANDEEP PABOLU
Class ID: 19

I. N-Gram

Ans:

1) Probability of sentence "I like green eggs and ham" using the appropriate bigram probabilities:

P(I | <s>) = 2/3 = 0.67

P(like | <s>) = 1/3 = 0.33

P(green | like) = 1/3 = 0.33

P(eggs | green) = 1/3 = 0.33

P(and | eggs) = 1/3 = 0.33

P(ham | and) = 1/3 = 0.33

P(</s> | ham) = 1/3 = 0.33

P(am | </s>) = 1/3 = 0.33

P(sam | </s>) = 1/3 = 0.33

2) Probability of sentence "I like green eggs and ham" using the appropriate Trigram probabilities:

P(I | <s> | like) = 1/3 = 0.33

P(like | I |green) = 1/3 = 0.33

P(green | like | eggs) = 1/3 = 0.33

P(eggs | green | and) = 1/3 = 0.33

P(and | eggs | ham) = 1/3 = 0.33

P(ham | eggs | </s>) = 1/3 =0.33

P(</s> | and | ham) = 1/3 = 0.33

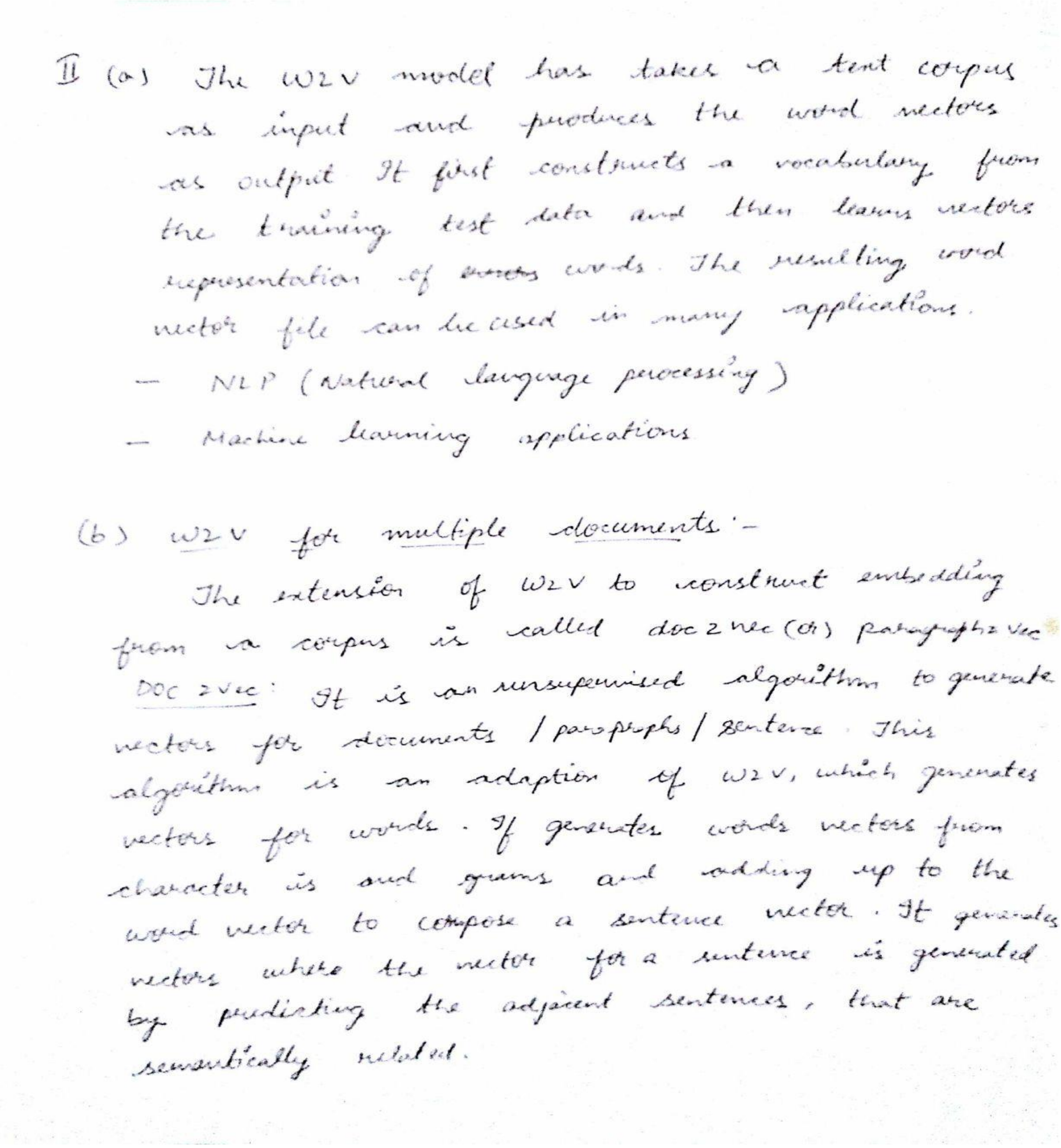
P(ham | </s> | and) = 1/3 = 0.33

II. Word2Vec

Answer:

a. Word2vec model:

Word2vec was created by a team of researchers led by Tomas Mikolov at Google. The algorithm has been subsequently analysed and explained by other researchers. Embedding vectors created using the Word2vec algorithm have many advantages compared to earlier algorithms like Latent Semantic Analysis.



II (a) The W2V model has taken a text corpus as input and produces the word vectors as output It first constructs a vocabulary from the training test data and then learns vectors representation of words. The resulting word vector file can be used in many applications.

— NLP (Natural language processing)

— Machine learning applications

(b) W2V for multiple documents :-

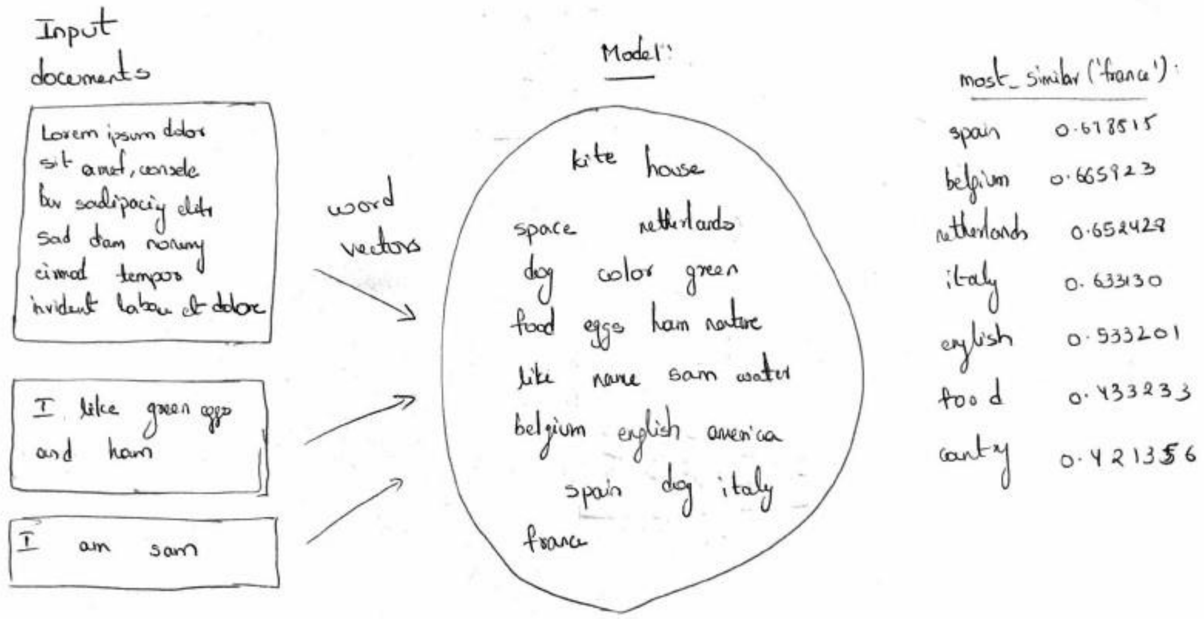
The extension of W2V to construct embedding from a corpus is called doc2vec (or) paragraph2vec

DOC 2Vec : It is an unsupervised algorithm to generate vectors for documents / paragraphs / sentence. This algorithm is an adaption of W2V, which generates vectors for words. If generates words vectors from character is and grams and adding up to the word vector to compose a sentence vector. It generates vectors where the vector for a sentence is generated by predicting the adjacent sentences, that are semantically related.

b. **Describe How to extend this model for multiple documents. Also draw a similar diagram for the extended model.**

The word2vec model can be extended for multiple documents by doc2vec. Doc2vec is an unsupervised algorithm to generate vectors for sentence/paragraphs/documents. [1405.4053] Distributed Representations of Sentences and Documents .

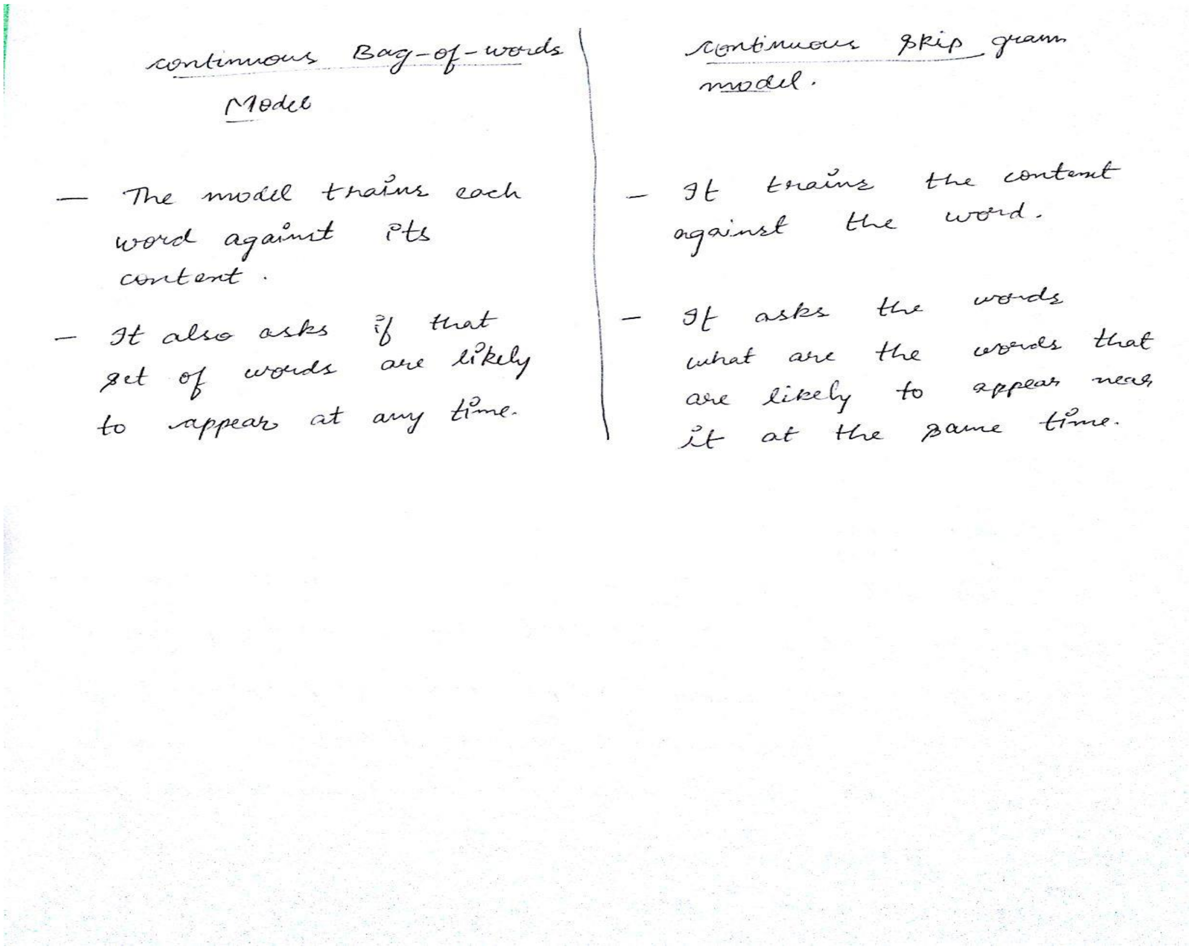
All the methods mentioned above are unsupervised algorithms requiring no training data.



Describe the differences of the following approaches

- Continuous Bag-of-Words model,
- Continuous Skip-gram model

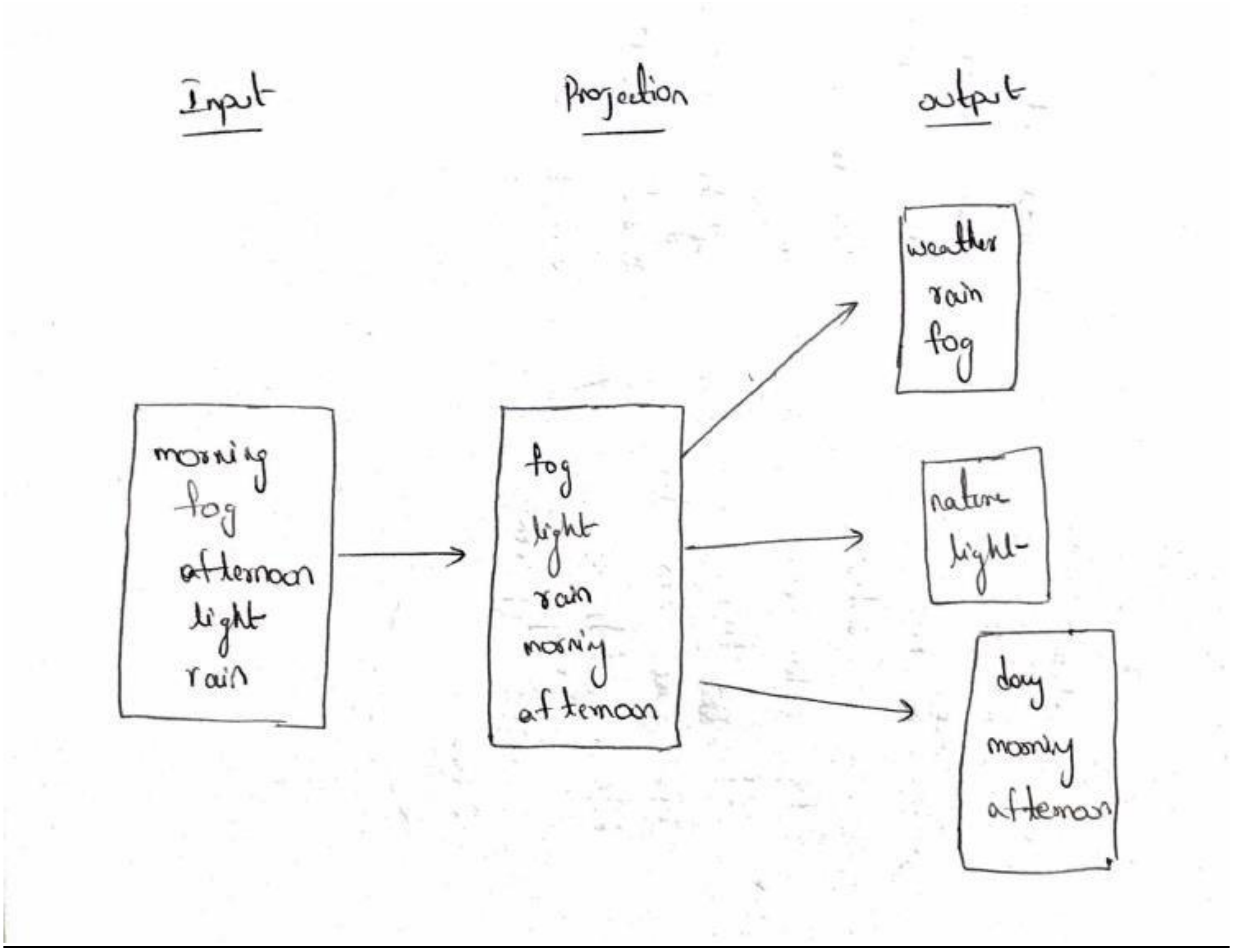
Answer:



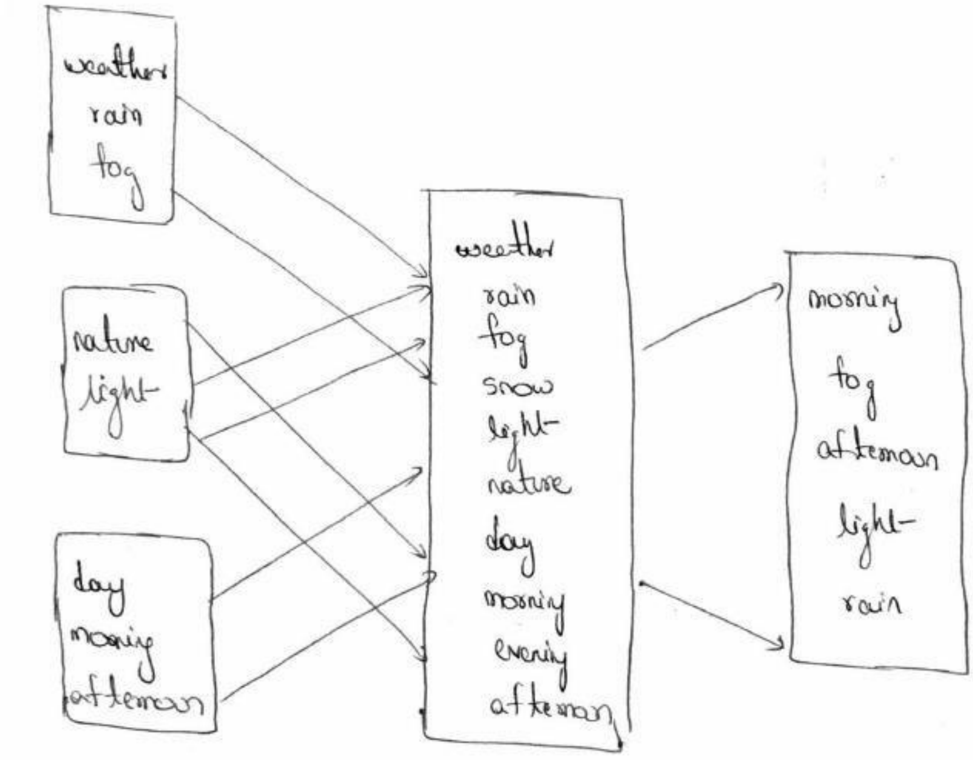
continuous Bag-of-words Model	continuous skip gram model.
— The model trains each word against its content.	— It trains the content against the word.
— It also asks if that set of words are likely to appear at any time.	— It asks the words what are the words that are likely to appear near it at the same time.

Answer: **skip-gram**

Word2Vec model:



CBOW model:


weather
rain
fog
nature
light
day
morning
afternoon
weather
rain
fog
snow
light
nature
day
morning
evening
afternoon
morning
fog
afternoon
light
rain