CS5560 Knowledge Discovery and Management

Problem Set 4 **Name: Sai Teja Makani**
June 26 (T), 2017 **Class ID: 12**

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:

CS5560 Knowledge Discovery and Management

Problem Set 4 **Name: Sai Teja Makani**
June 26 (T), 2017 **Class ID: 12**

a. Word2vec model:

Word2vec is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.

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

The w2v model has taken a text corpis as input and produces the word referes as output it first constructs a vocabulary from the trimming test and then lemma vectors representations of words in manu applications.

- ➔ NLP (nalutal Language Programming)
- ➔ Machine Learning Applications.

b. W2V for Multiple Documents:

The extension of w2v to construct embodied from a corpus is calls doc2V. It is an unactioned algorithm to generate vectors for documents/ sentences. This algorithm is am adaption of w2v, which generates vectors for words. If generates words up to the word vectors from character is and from and adding up t0 the word vector to compare a sentence vector. It generates vectors where the vector for a sentence is generated by predicting the adjacent sentences, that are by predicting the adjacent sentences, That are semantically reversal.

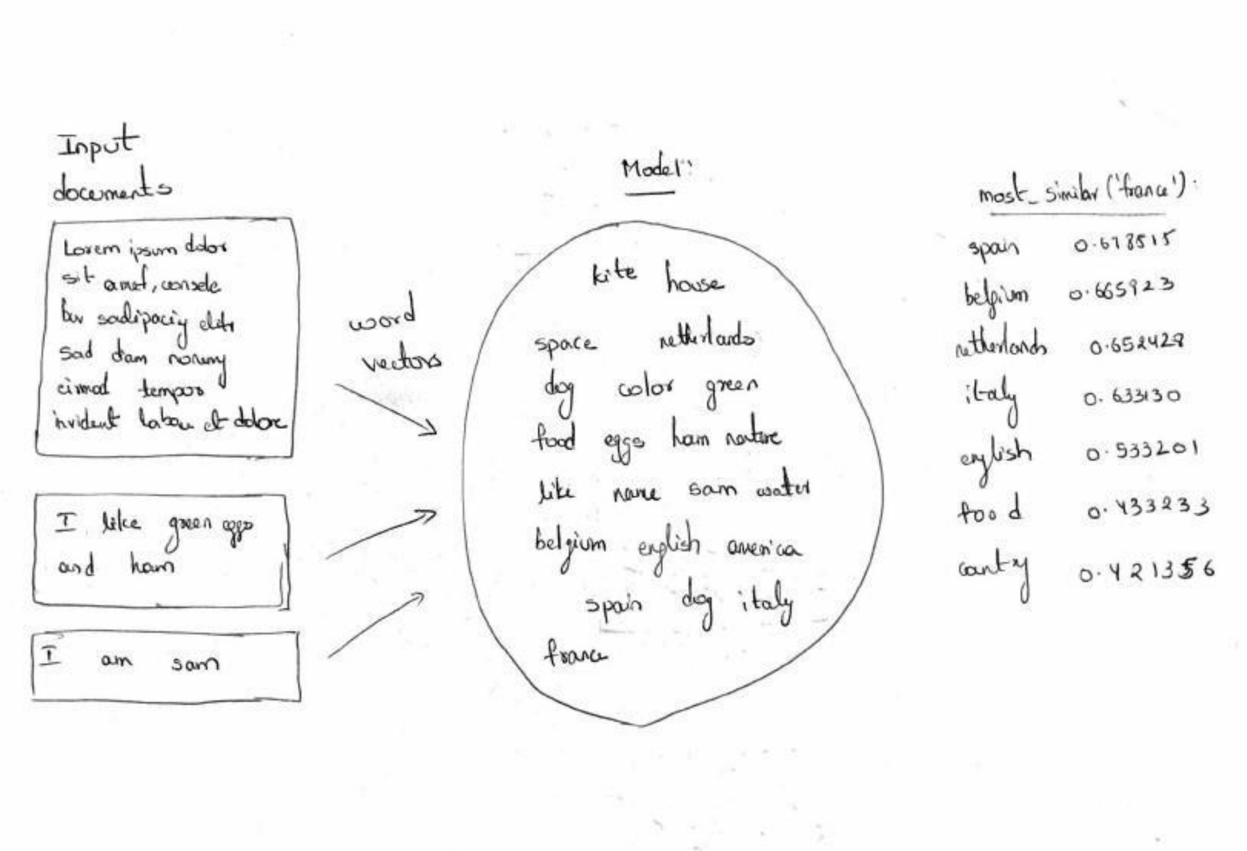
The word2vec model can be extended for multiple documents by doc2vec. Doc2vec is an unsupervised algorithm to generate vectors for sentence/paragraphs/documents. Distributed Representations of Sentences and Documents .

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

CS5560 Knowledge Discovery and Management

Problem Set 4 **Name: Sai Teja Makani**

June 26 (T), 2017 **Class ID: 12**



most_similar('france'):

spain	0.678515
belgium	0.665923
netherlands	0.652429
italy	0.633130
english	0.533201
food	0.433233
country	0.421356

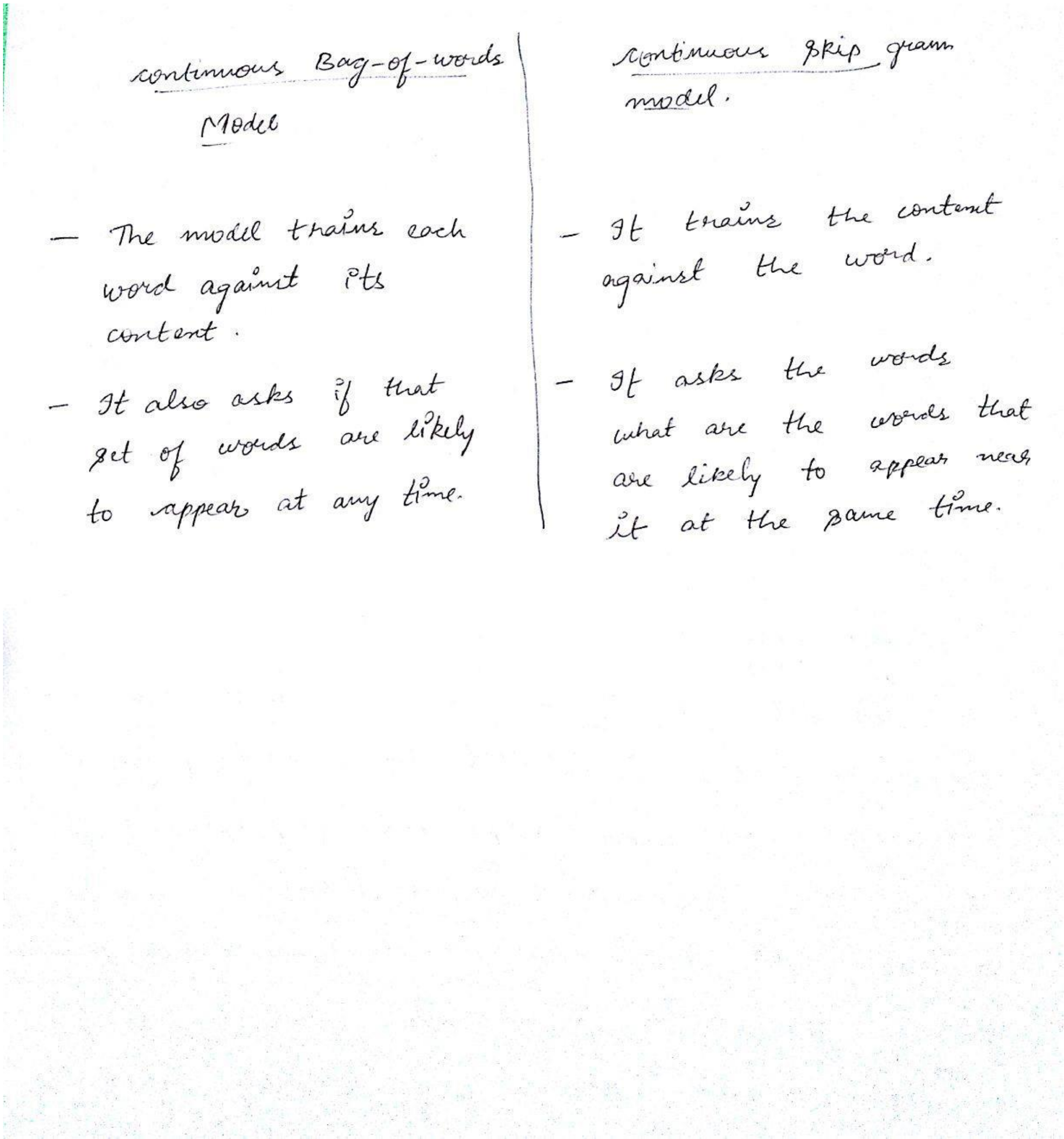
Describe the differences of the following approaches

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

Answer:

CS5560 Knowledge Discovery and Management

Problem Set 4 **Name: Sai Teja Makani**
June 26 (T), 2017 **Class ID: 12**



continuous Bag-of-words Model

- The model trains each word against its content.
- It also asks if that set of words are likely to appear at any time.

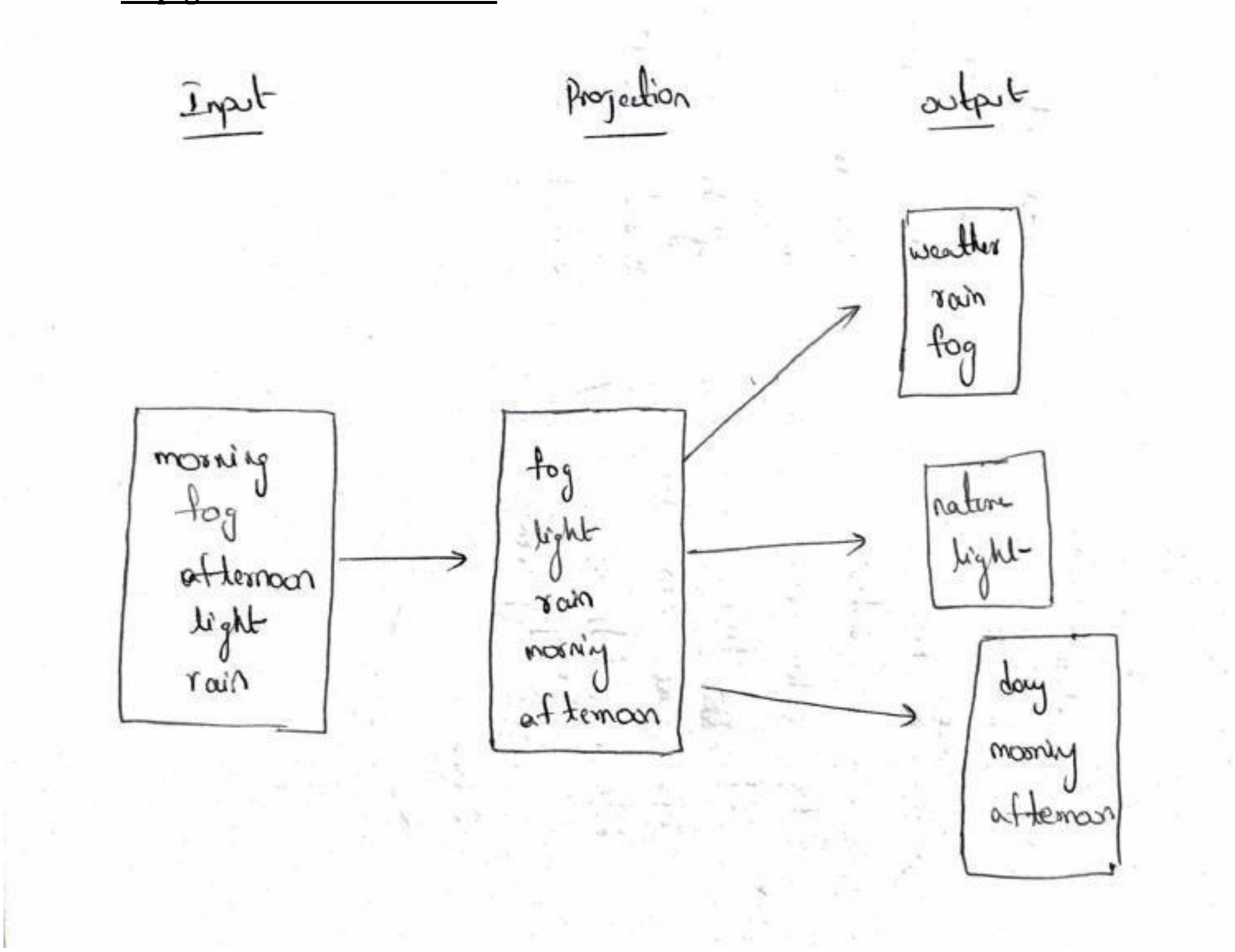
continuous skip gram model.

- It trains the content against the word.
- It asks the words what are the words that are likely to appear near it at the same time.

CS5560 Knowledge Discovery and Management

Problem Set 4 **Name: Sai Teja Makani**
June 26 (T), 2017 **Class ID: 12**

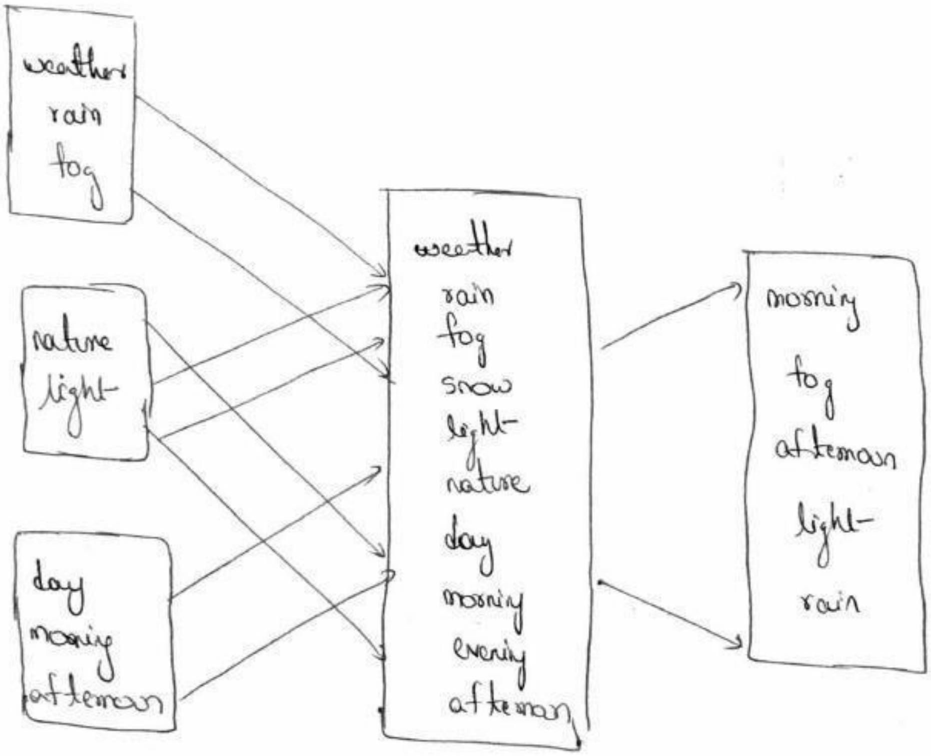
Answer: skip-gram Word2Vec model:



CBOW model: Word2vec can utilize either of two model architectures to produce a distributed representation of words: continuous bag-of-words (CBOW) or continuous skip-gram. In the continuous bag-of-words architecture, the model predicts the current word from a window of surrounding context words. The order of context words does not influence prediction (bag-of-words assumption). In the continuous skip-gram architecture, the model uses the current word to predict the surrounding window of context words. The skip-gram architecture weighs nearby context words more heavily than more distant context words.According to the authors' note, CBOW is faster while skip-gram is slower but does a better job for infrequent words.

CS5560 Knowledge Discovery and Management

Problem Set 4 **Name: Sai Teja Makani**
June 26 (T), 2017 **Class ID: 12**


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