PGR 210 - Natural Language Processing Part

Kristiania University College

By Huamin Ren

Huamin.ren@kristiania.no

Høyskolen Kristiania

Outline of week 45

1. GloVe
2. Similary metric
3. K-means clustering

1. GloVe

Høyskolen Kristiania

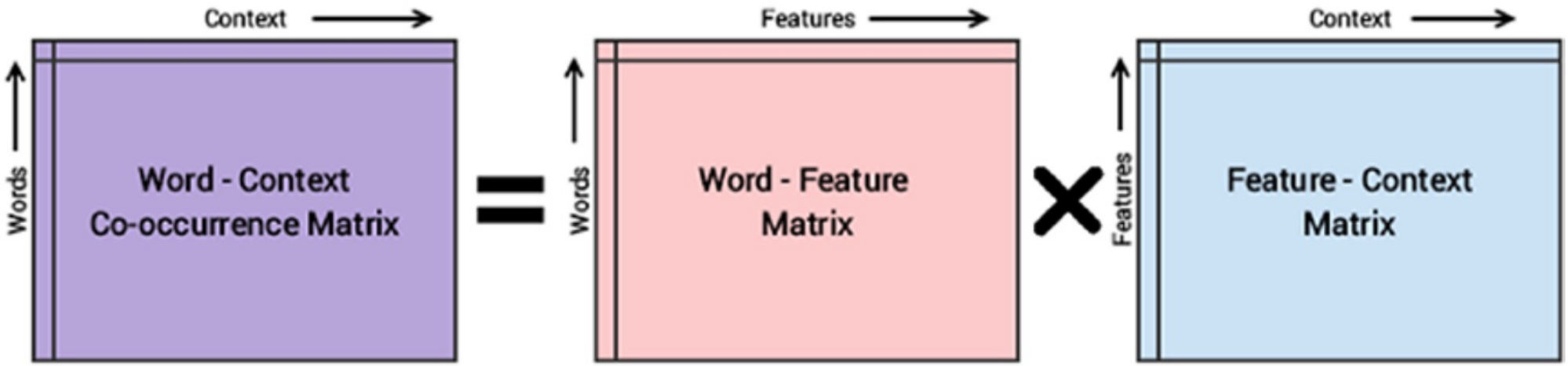
1.1

- Word2vec was a breakthrough, but it relies on a neural network model that must be trained using backpropagation.
- GloVec is applying direct optimization of the global vectors of word co-occurrences (co-occurrences across the entire corpus)

https://nlp.stanford.edu/pubs/glove.pdf?fileGuid=WyYwxqq8kWjKdWgd

Høyskolen Kristiania

1.2 Conceptual model for the GloVe model's implementation



Høyskolen Kristiania

1.3

- Considering the Word-Context (*WC*) matrix, Word-Feature (*WF*) matrix, and Feature-Context (*FC*) matrix, we try to factorize

$$\boldsymbol{WC} = \boldsymbol{WF}\ \boldsymbol{FC}$$

- Typically initialize *WF* and *FC* with some random weights and attempt to multiply them to get *WC'* (an approximation of WC) and measure how close it is to *WC*.
- Do this multiple times using *Stochastic Gradient Descent* (SGD) to minimize the error.

Høyskolen Kristiania

- WF (*Word-Feature):* word embeddings for each word, where *F* can be present to a specific number of dimensions.
- Similarity between Word2Vec and: build a vector space where the position of each word is influenced by its neighboring words based on their context and semantics.
- Difference in the two models: Word2Vec starts with local individual examples of word co-occurrence pairs; GloVe starts with global aggregated co-occurrence statistics across all words in the corpus.

Høyskolen Kristiania

2. Similarity metric

Høyskolen Kristiania

2.1 Similarity scores

- Similarity scores (and distances) tell on how similar or how far apart two documents are based on the similarity (or distance) of the vectors you used to represent them
- Some familiar distances:
  - Euclidean or Cartesian distance, or root mean square error (RMSE): 2-norm or L2 , MSE,
  - Squared Euclidean distance, sum of squares distance (SSD): L2
  - Cosine or angular or projected distance: normalized dot product
  - Minkowski distance: p-norm or Lp
  - Fractional distance, fractional norm: p-norm or Lp for $0 < p < 1$
  - City block, Manhattan, or taxicab distance; sum of absolute distance (SAD)

- V1, V2
- Sim(V1, V2) = 0.99999999
- Sim(V1, V1) = 1, 1 m documents, v1, v2, v1m
- Distance(V1, V1) = 0
- Sim(V1, V2) =0.49
- Sim(V1, V3) = 0.111
- Sim(V1, V4) = 0.47
- Threshold=0.48

Pairwise distances available in sklearn

- 'cityblock', 'cosine', 'euclidean', 'l1', 'l2', 'manhattan', 'braycurtis', 'canberra', 'chebyshev', 'correlation', 'dice', 'hamming', 'jaccard', 'kulsinski', 'mahalanobis', 'matching', 'minkowski', 'rogerstanimoto', 'russellrao', 'seuclidean', 'sokalmichener', 'sokalsneath', 'sqeuclidean', 'yule'

2.2 Distance

- Distance measures are often computed from similarity measures (scores) and vice versa such that distances are inversely proportional to similarity scores.
- For distances and similarity scores that range between 0 and 1, like probabilities, it's more common to use a formula like this:

  similarity = 1. - distance

  distance = 1. - similarity

Høyskolen Kristiania

$$hd(u,v) = \sum_{i=1}^{n} (u_i \neq v_i)$$

$$norm\_hd(u,v) = \frac{\sum_{i=1}^{n} (u_i \neq v_i)}{n}$$

U=[1 1 1 0 0] v=[0 0 0 1 1]
What is hd? What is norm_hd?

Høyskolen Kristiania

2.3 Use similarity on?

- **Lexical similarity:** This involves observing the contents of the text documents with regards to its syntax, structure, and content and measuring their similarity based on these parameters.
- **Semantic similarity:** This involves determining the semantics, meaning, and context of the documents and then determining how close they are to each other.

Høyskolen Kristiania

- **Term similarity:** Similarity between individual tokens or words
- **Document similarity:** Similarity between entire text documents

Høyskolen Kristiania

3. K-means Clustering

Høyskolen Kristiania

- The k-means clustering algorithm is a centroid-based clustering model that tries to cluster data into groups or clusters of equal variance.
- Disadvantage of this algorithm is that the number of clusters (**k**) needs to be specified in advance.
- Advantage: perhaps the most popular clustering algorithm, due to its ease of use as well as it being scalable with large amounts of data.

Høyskolen Kristiania

Mathematical definition

- A dataset **X** with **N** data points or samples, the task is to group them into K clusters, where **K** is a user-specified parameter.
- The k-means clustering algorithm will segregate the **N** data points into **K** disjoint separate clusters, C_k, and each of these clusters can be described by the means of the cluster samples.

X = {x1,x2…xn}
c1 = {x1, x2} c2={x3}…c5…
X1=[1 0 0 1 1 1] x2=[1 0 0 1 1 0] c1 = [1 0 0 1 1 0.5]
c1={mean(x1,x2)}

- These means become the cluster centroids μ_k such that these centroids are not bound by the condition that they have to be actual data points from the **N** samples in **X**.

Høyskolen Kristiania

K-means procedure

Steps:

1. Choose initial **k** centroids μk by taking **k** random samples from the dataset **X**.
2. Update clusters by assigning each data point or sample to its nearest centroid point.
3. Recalculate and update clusters based on the new cluster data points for each cluster obtained from Step 2. Mathematically this can be represented as follows:

1. k=5
U1(x2)….u5(x1)
2. For each data xi:
Compute distances,
Find the closest cluster
Assign xi to u5
3. recalculate:
u1:=(x2)
U2 =(x3)

$$C_k = \{x_n : \|x_n - \mu_k\| \leq all \|x_n - \mu_l\|\}$$

$$\min \sum_{i=1}^{K} \sum_{x_n \in C_i} \| x_n - \mu_i \|^2$$

Høyskolen Kristiania