



Item2Vec: Neural Item Embedding for Collaborative Filtering

Abstract

많은 CF 알고리즘들은 아이템-기반이라는 의미를 가진다. 아이템-아이템의 관계를 분석하는 것이다. 아이템 유사도를 생성하기 위해서. 최근 자연어 처리 필드에서의 업적들은 제안한다. 뉴럴 임베딩 알고리즘을 사용해서 words의 latent representation을 학습을 제안한다. SGNS(skip-gram with negative sampling) word2vec이라고 알려진, linguistic task들에서 SOTA를 보였다. 본 논문에서는 아이템-기반의 CF가 뉴럴 임베딩과 동일한 프레임워크라는 것을 보인다. SGNS에서 영감을 받아, 유저 프로파일 정보가 없어도 item-item의 관계를 추론하는 뉴럴 임베딩을 보인다. 그리고 성능을 SVD와 비교했을 때 경쟁력을 가진다는 것을 보인다.

Introduction

싱글 아이템 추천은 전통적인 user-to-item 추천과 조금 더 다르다. 그 이유는 특정 아이템에 대한 유저들의 interest와 유저가 구매하려는 의도를 보이기 때문이다. 그래서 싱글 아이템 추천들은 유사도 종종 높은 CTR을 기반으로 한다. user-to-item 추천들의 결과적으로 책임감 있는 더 큰 판매나 수익보다. 아이템 유사도들을 기반으로한 싱글 아이템 추천들은 다양한 추천 태스크들에 사용되어져 왔다. user-item CF 방법이 바로 item 간 관계를 학습하는 것보다 더 나은 item representation을 생성했다.

본 논문에서는 SGNS를 item 기반의 CF에 적용하는 것을 제안한다.

Relative Works

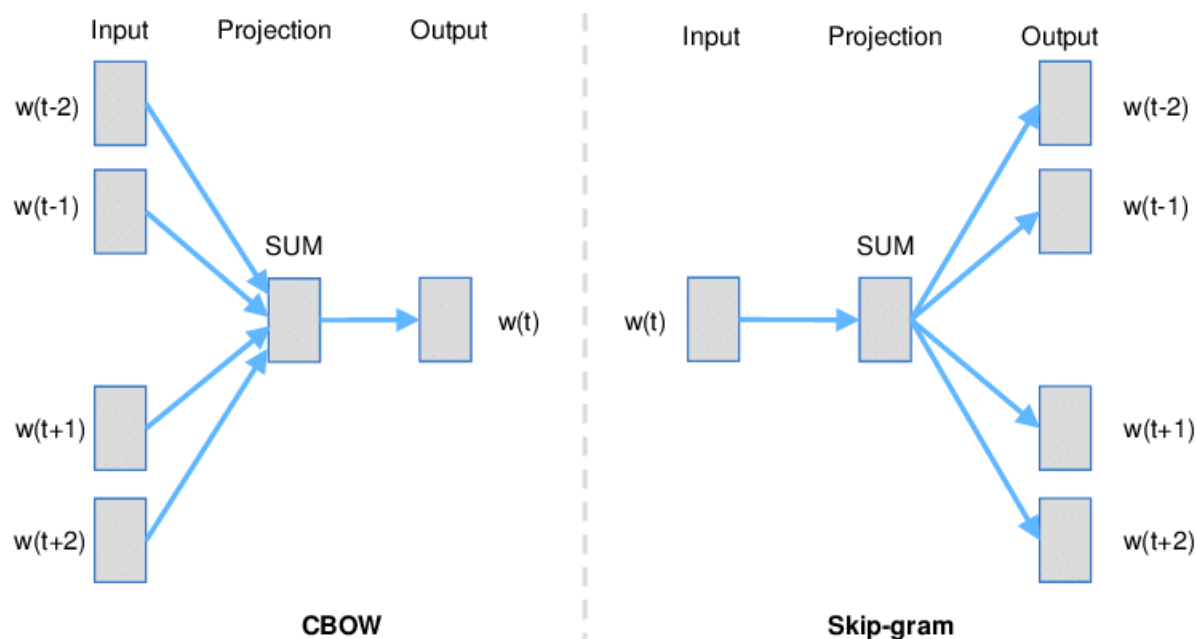
SGNS에 대한 공식과 식을 설명하고 있음

Item2Vec - SGNS for item-based CF

skip-gram, cbow, negative sampling

CBOW : 주변에 있는 문맥 단어 (context word)들을 가지고 타깃 단어 하나를 맞추는 과정으로 학습

SKIP-GRAM : 타깃 단어를 가지고 주변 문맥 단어가 무엇일지 예측하는 과정으로 학습




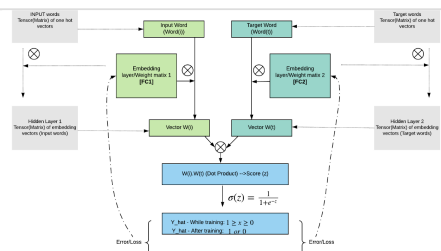
Positive Sample : 타깃 단어(t)와 그 주변에 실제로 등장한 문맥 단어(c) 쌍을 가리킨다

Negative Sample : 타깃 단어와 그 주변에 등장하지 않는 단어 (말뭉치 전체에서 랜덤 추출) 쌍을 의미

Word2Vec -Negative Sampling made easy

This is my second post on Word2Vec. The previous article was about the probabilistic model explaining the mechanics of embedding and appropriately using vector representation. You

 <https://medium.com/towardsdatascience/word2vec-negative-sampling-made-easy-7a1a647e07a4>



타겟(Target) 단어와 문맥(Context) 단어 쌍이 주어졌을 때 문맥(Context) 단어가 무엇일지 맞추는 이진 분류 과정에서 학습된다.

Negative Sampling

학습 시에 1개의 포지티브 샘플과 k개의 네거티브 샘플만 계산한다.

$$P(w_i) = \frac{f(w_i)}{\sum_{j=0}^n (f(w_j))}$$

$$P(w_i) = \frac{f(w_i)^{3/4}}{\sum_{j=0}^n (f(w_j)^{3/4})}$$

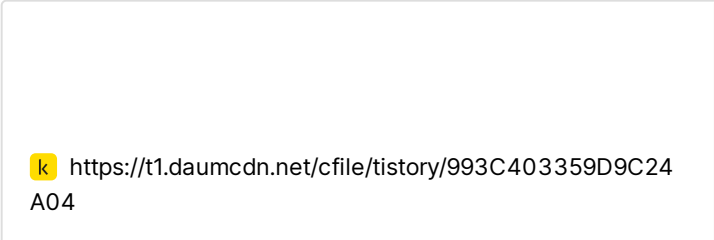
말뭉치에 자주 등장하지 않는 희귀한 단어가 네거티브 샘플로 조금 더 잘 뽑힐 수 있도록 설계했다.

Sub Sampling

Skip-Gram은 많은 학습 데이터 쌍을 만들 수 있기 때문에 고빈도 단어의 경우 등장 횟수만큼 모두 학습 시키는 것이 비효율적이다. 학습량을 효과적으로 줄여 계산량을 감소시키는 전략이다.

<https://t1.daumcdn.net/cfile/tistory/993C403359D9C24A04>

고빈도 단어의 경우 등장 횟수만큼 모두 학습시키는 것이 비효율적이다



k <https://t1.daumcdn.net/cfile/tistory/993C403359D9C24A04>

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$$

수식 안의 t가 한계점이다.
 w_i 가 잘 나타낼수록 $P(w_i)$ 값이 크다.
 $P(w_i) > t$ 이면, w_i 는 subsampling을 한다.

등장 비율이 적은 단어는 나올 때마다 빼놓지 않고 학습을 진행한다.

Word2Vec 그리고 추천시스템의 Item2Vec

Recommender System with Distributed Representation

Published on In recent years, Word2Vec and its expansion (Doc2Vec, Paragraph2Vec, etc.) is receiving a lot of attention in the NLP field. In this slide, we will introduce our approach for

<https://www.slideshare.net/rakutentech/recommender-system-with-distributed-representation>

分散表現を用いた
商品レコメンダーシステムの構築と評価
Recommender System with Distributed Representation

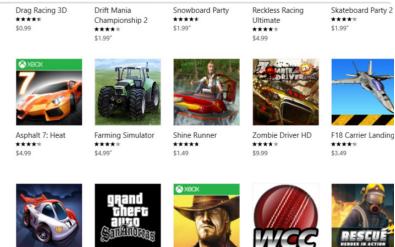
Thuy Phivan^{1,2}, Chen Liu² and Yu Hirata²

1. Computational Linguistics Laboratory, NAIST
2. Rakuten Institute of Technology, Rakuten, Inc.
{ar-thuy.phivan, chen.liu, yu.hirate}@rakuten.com

Item2Vec: Neural Item Embedding for Collaborative Filtering

많은 협업 필터링 (CF, Collaborative Filtering) 알고리즘은 아이템 유사도를 생성하기 위해 아이템-아이템 간 관계를 분석한다는 점에서 아이템 기반이라고 할 수 있음. 최근 자연어 처리 분야에서의 연구들 중 일부는 뉴럴 임베딩 알고리즘

https://soobarkbar.tistory.com/129



item2vec/kako buffalo

가정

- 아이템 임베딩 (4000, 300)
- 브랜드 임베딩 (20, 300)
- 더미 임베딩 (300, 브랜드 임베딩)

임베딩 수 → 공식 찾아보기

아이템 임베딩 + 브랜드 임베딩

브랜드임베딩 x 아이템 임베딩 → (20 x 4000)

→ 아이템-브랜드 임베딩 (4000, 20)

→ 아이템 -아이템 임베딩 (4000, 4000)

상품 명에 대한 임베딩을 뽑아서 상품 명 간의 유사도로 추천을 하겠다.

브랜드 상세 연관 상품 추천 (이 브랜드를 추천)

1. Question

아이템 코드 임베딩 + 브랜드 임베딩 ?

브랜드 임베딩

- 브랜드 필터링 필수

ex) 삼성 제품을 본다면 삼성 브랜드 상품을 추천

2. EDA

조회 7일, 구매 30일, 장바구니

클릭기반, 윈도우 크기 2기

3. Model Mapping (Item2Vec)

평가 지표 : 아이템 들 간의 평균 거리

: 평가 지표

: 전환 평가

4. Developer

5. Delivery