Проект по “Препоръчващи системи”

Работа по състезанието RecSys Challenge 2022

Иван Арабаджийски

Резюме

Курсовият проект се състои върху работа по съсзението “RecSys Challenge 2022”. Състезанието се организира от Dressipi - водеща компания за препоръчващи системи в областта на модата. Тази година задачата в състезанието се фокусира върху препоръчване на модни артикули. По дадени потребителски сесии, данни за покупките и характеристики на продуктите трябва да се предскаже кой продукт ще се закупи накрая на сесията. Характеристиките на продуктите са много и разнообразни - цвят, дължина, ръкави, стил и тн. Характеристиките са анотирани ръчно от специалисти на Dressipi, проверени за коректност и коригирани при нужда, така че се очаква данните да са с високо качество. Като избран подход за решаване на тази задача, този курсов проект избира метод за препоръчваща система базирана на сесии, който се базира на идеята за ембединги. Представяйки сесиите като поредица от разгледаните продукти образуваме ембдинги чрез алгоритъпа prod2vec, който взаимства идеята си от подобен алгоритъм в средата на обработката на естествен език - word2vec и по-конкретно skip-gram. Там по дадена дума се предсказва нейния контекст. Идеята на алгоритъма prod2vec е същата, по подадена сесия от продукти, след образуването на ембедингите, се опитваме да предскажем най-близките до нея тродукти, т.е. тези, които е най-вероятно да бъдат купени.

Въведение

С напредване на времето нуждата от препоръчващи системи става все по-голяма и по-голяма. Всички магазини, платформи за филми, музика и други вече разчитат на събирането на потребителски данни като кликове, време прекарано на дадена страница и купуване на продукти, за да изградят една по-добра препоръчваща система. Това е важно за общото добро впечатление на клиентите и води до повече печалби. Естествено, да се предскажат продукти, които биха предизвикали интереса на даден потребител не е лесна задача. Във времето са се появили различни методи за разработка на препоръчващи системи, но през последните години невронните мрежи и по-специално дълбоките невронни мрежи са се утвърдили като метод за решаване на такива задачи. В задачата се изисква по дадена сесия да се предскажат следващите 100 най-вероятни да бъдат закупени продукти от конкретния човек подредени по ранг. Дадени са тренировъчни сесии, купените продукти за всяка тренировъчна сесия, характеристиките на всеки продукт, и продуктите, измежду които трябва да избираме. След това имаме предоставени и

тестовите сесии, за които трябва да направим предсказанията. За всяка тренировъчна сесия имаме и продукт, който потребителя всъщност наистина е купил. Този продукт наричаме ground truth. Оценката се извършва чрез функцията MRR (Mean Reciprocal Rank), която дава по-високи резултати, ако истинският продукт е по-напред в класацията.

Преглед на областта

В началото препоръчващите системи са били непорсонализирани, базирани на съдържание. Препоръките са били правени на базата на оценките, които потребителите дават на конкретен продукт. Естествено Това води до много проблеми. Продуктите с много положителни оценки трудно се свалят от челните позиции, а тези без никакъв рейтинг не могат да стигнат дотам да бъдат препоръчане и съответно остават без рейтинг. След време се преминава към колаборативни препоръчващи системи. Те разчитат на общите интереси между потребителите и за това препоръчват подобни продукти на потребители с подобни интереси. Тук се разглежда не толкова качеството на продукта сам по себе си, колкото това, че ако той се е харесал на потребител с подобни на нашите интереси, то е силно вероятно да се хареса и на нас. Хибридните препоръчващи системи използват резултати от двата основни вида споменати по-горе като използват отегляване или просто осредняване между различните си компоненти. С появата на нови технологии като дълбоките невронни мрежи и нови модели базирани на тях като трансформърите започват да се появяват и нови идеи за препоръчващи системи. Появяват се адаптации на много от отвърдените модели вече за препоръчване - transformers4rec, bert4rec и тн.

Данни и характеристики

Данните са предоставени от организаторите на състезанието. Те представляват 1.1 милиона сесии на истински потребители, които завършват с покупка. Също всички продукти са описани чрез характеристики, които са също предоставени. Предоставените данни са анонимни. Сесиите представят всички разгледани продукти от даден потребител, всяка сесия е в рамките на един ден, т.е. Разглежда се активността на потребител за ден. Покупка се случва накрая на всяка сесия. Имаме по една покупка за сесия. Характеристиките на продуктите във формат “цвят: зелен, деколте: изрязано”. Предоставени са четири файла за трениране и един тестови. Тренировъчни сесии, покупките, на края на всяка от тренировъчните сесии, характеристиките на продуктите и продуктите, от които да предсказваме. От всички данни организаторите на състезанието са избрали 1 милион сесии за тренировъчни, 50 хиляди за тестови и 50 хиляди за финален тест. Дадените сесии са сесии, които завършват с поне една покупка. При изработването на данните от сесии с повече от една покупка е избрана произволно, коя да се вземе като покупка за сесията. След това сесията се отрязва до момента на първото срещане на покупката (не включително) и се определя като валидна сесия. Причината за това е да се разгледат потребители в различни моменти от тяхната сесия. Целта е да се предскажат продуктите, които те биха си купили максимално рано. Метриката за оценка на резултатите е MRR (Mean Reciprocal Rank). Колкото по-нагоре в ранкинга е купеният продукт толкова по-висока е оценката. Това е метрика в статистиката, която се използва при оценка на процес, който извежда списък от възможни резултати, подредени по ранг впрямо верността им.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

Например

Query	Proposed Results	Correct response	Rank	Reciprocal rank
cat	Catten, cati, cats	cats	3	1/3
torus	Torii, tori, toruses	tori	2	1/2
virus	Viruses, virii, viri	viruses	1	1

Подадените данни съдържат един милион сесии за обучение, средно всяка сесия има по 4.73 продукта. Данните са изчистени и няма нулеви полета никъде. Уникалните продукти от обучаващите са на брой. Продуктите имат между 2 и 33 характеристики.

Методи

За решаването на този проблем ще използваме препоръки, базирани на сесиите. Най-простият и често срещан метод е препоръчването на продукта, който най-често се среща заедно с последния продукт от сесията. Този подход обаче не може да улови достатъчно добре цялостната информация, която се съдържа в сесията, а повече разчита на продукт-продукт препоръката. Ние ще подходим по различен начин. Ще третираме проблема като проблем от областта на обработката на естествен език и ще приложим алгоритъма word2vec. Това е проста невронна мрежа с един скрит слой, който обучава ембединги за всяка дума, запечатвайки семантичното ѝ значение. В случая на алгоритъпа skip-grap имаме подадена дума, по която предсказваме нейния контекст. Как можем да използваме тази идея за препоръчваща система на базата на сесии? Ще представим всяка сесия като поредица от продуктите, които сме разгледали, по този начин всеки продукт ще бъде нашата “дума” в контекста на word2vec. Всичките сесии ще бъдат нашия корпус. Ще научим ембдеинги, които да запечатат отношенията между продуктите в смисъла на сесиите и срещанията им един след друг. Тези ембединги ще съдържат повече информация спрямо някоя проста евристика и са по-бързи за пресмятане от повечето сложни и бавни алгоритми, които изискват много данни. Всъщност представянето на този проблем по гореспоменатият начин има смисъл, защото имаме последователни данни и подредбата им е вероятно да няма значение. В избрания подход все пак ги сортираме по дата.

Сценарият за построяване на ембедингите, тестване, валидиране и предсказване е следният. Първо представяме всяка сесия като масив от идентификационните номера на продуктите видени от потребителя в нея, сортирани по час. Използваме word2vec, за да построим ембедингите. Хиперпараметрите имат значение. В случая са използвани 5 епохи с алгоритъма skip-gram и прозорец голям колкото най-дългата сесия. След това имаме валидационна сесия с дължина n. Искаме да предскажем следващите 100

продукта подредени спрямо вероятността те да бъдат купени. За валидационните сесий знаем истинският продукт, който е бил купен. Него наричаме истина (ground truth). Използваме го след това, за да оценим получените предсказания. За предсказване на дадените тестови сесии взимаме стоте най-близки продукта, използвайки косинусова близост.

Резултати

Получените резултати се изчисляват автоматично от системата на състезанието. Като финален резултат за състезанието резултатът е 0.07919231889694121. Получени резултати върху валидационно множество са в порядъка на 0.11.

Литература

Choi, M. Kim, J. Lee, J. (2021, June 25th). *Session-aware Linear Item-Item Models for Session-based Recommendation*. Retrieved July 3, 2022, from https://arxiv.org/pdf/2103.16104v2.pdf

Cloudera. (2021, May). *Session-based Recommender Systems*. Session-based Recommenders. Retrieved July 2, 2022, from https://session-based-recommenders.fastforwardlabs.com/

How to build a session-based recommender using word2vec. RecoStep. Retrieved July 4, 2022, from https://step.recohut.com/codelabs/session-based-recommender-using-word2vec/index.html?index=..%2F..index#0

Landman, E. (2022, June 4th). *Session-Based Recommender Systems with Word2Vec*. Session-based Recommenders. Retrieved July 4, 2022, https://towardsdatascience.com/session-based-recommender-systems-with-word2vec-666afb775509

Vasile, F. Smirnova, E. Conneau, A. (2016, July 25th). *Meta-Prod2Vec - Product Embeddings Using Side-Information for Recommendation*. Retrieved July 2, 2022, from https://arxiv.org/pdf/1607.07326.pdf