

BAG OF WORDS VA WORD2VES MODELLARI TAQQOSLANISHI

Farg'ona Davlat Texnika Universitet
Zaynobiddinov Shohrux Xusniddin o'g'li
shohruxzaynobiddinov1@gmail.com
+998953530054

Annotatsiya

Ushbu maqolada tabiiy tilni qayta ishlash jarayonida matnni vektor ko'rinishiga keltirishning ikki asosiy yondashuvi — Bag of Words va Word2Vec modellari batafsil tahlil qilinadi. Bag of Words modeli so'zlarning chastotasiga asoslangan sodda statistik yondashuv bo'lib, matndagi semantik ma'noni aks ettirmaydi va lug'at hajmining kattaligi sababli o'lcham muammosiga ega. Word2Vec modeli esa neyron tarmoqlar asosida ishlaydi va so'zlar orasidagi kontekstual hamda semantik aloqalarni o'rganib, ma'noga ega zich vektorlar yaratadi. Annotatsiyada har ikkala modelning ishlash mexanizmi, afzallik va cheklovlari, amaliy qo'llanish imkoniyatlari hamda zamonaviy NLP tizimlarida ularning o'ri yoritilgan. Maqola ushbu modellarni tanlashda vazifa murakkabligi, resurs talabi va semantik chuqurlik kabi mezonlarning ahamiyatini ochib beradi.

Kalit so'zlar: Bag of Words, Word2Vec, NLP, vektorlashtirish, semantika, kontekst, CBOW, Skip-gram, TFIDF, neyron tarmoqlar.

Kirish. Tabiiy tilni qayta ishlash sohasida matnni raqamli ko'rinishga keltirish kompyuterlar tomonidan tilni tushunish jarayonining boshlang'ich bosqichidir. Har bir matnni sonlar ketma ketligi sifatida ifodalash zarur, chunki mashinalar faqat raqamlar bilan ishlay oladi. Matnni vektor shakliga keltirish uchun bir nechta yondashuvlar ishlab chiqilgan bo'lib, ularning ikki eng mashhuri Bag of Words va Word2Vec modellaridir. Bu ikki yondashuv maqsadi bir xil bo'lsa ham, ishlash prinsipi, semantik imkoniyatlari, samaradorligi va natijalari bo'yicha keskin farqlanadi.

Bag of Words modeli.

Bag of Words matnni so'zlar to'plami sifatida ko'ruvchi eng klassik yondashuvdir. Ushbu model matndagi so'zlarning tartibini hisobga olmaydi, balki faqat qaysi so'z necha marta uchraganiga e'tibor qaratadi. Masalan, bir nechta hujjatlardan iborat korpus berilsa, avvalo umumiy lug'at tuziladi va har bir hujjat shu lug'atdagi so'zlarning chastotasi asosida vektorga aylantiriladi. Agar lug'atda o'n minglab so'zlar bo'lsa, har bir matn shu uzunlikdagi vektordan iborat bo'ladi. Bu yondashuv matematik jihatdan juda sodda, lekin semantik ma'no, kontekst va tilning ichki tuzilmasini aks ettira olmaydi. Bag of Wordsning ikki keng qo'llaniladigan varianti mavjud. Birinchisi, oddiy chastotani hisoblaydigan Count Vectorizer bo'lib,

unda soʻz qanchalik koʻp uchrasa, uning vazni shunchalik yuqori boʻladi. Ikkinchisi, TF-IDF yondashuvi boʻlib, unda kam uchraydigan, lekin mazmunli soʻzlar yuqori ahamiyatga ega deb hisoblanadi. TF-IDF koʻpincha katta matnlarda sifatni oshiradi, chunki umumiy va keng tarqalgan soʻzlarning modeli chalgʻitishi kamayadi.

Bag of Words modeli afzalliklari jihatidan sodda, tez hisoblash imkonini beradi va koʻplab klassik mashinaviy oʻrganish algoritmlari bilan juda mos keladi. Kamchiliklari esa ancha jiddiy. Model semantik maʼnoni tushunmaydi, sinonimlarni farqlay olmaydi, soʻzlarning tartibini inkor qiladi va lugʻat hajmi ortgani sayin vektorlar juda katta oʻlchamli boʻlib boradi. Shuning uchun mazmunni chuqur anglash talab qilinadigan vazifalarda Bag of Words yetarli boʻlmaydi.

Word2Vec modeli

Word2Vec matnni vektorlashning zamonaviy yondashuvidir. U kichik neyron tarmoq asosida soʻzlarni ularning maʼnosiga qarab zich vektorlarga aylantiradi. Ushbu model soʻzlarning kontekstini oʻrganadi, yaʼni har bir soʻz yon atrofidagi soʻzlar bilan qanday munosabatda ekanligini tahlil qiladi. Natijada har bir soʻz uchun maʼnoli, semantik jihatdan boy vektor yaratiladi. Bunday vektorlar oʻzaro yaqin maʼnoli soʻzlarni bir biriga yaqin joylashtiradi, sinonimlar bir xil makonda toʻplanadi, hatto matematik munosabatlar ham shakllanadi. Word2Vec ikki asosiy arxitekturaga ega. Birinchisi CBOW boʻlib, u kontekstdan soʻzni bashorat qiladi va katta matnlarda juda tez ishlaydi. Ikkinchisi Skip-gram boʻlib, u soʻzdan kontekstni bashorat qiladi va kam uchraydigan soʻzlarni oʻrgatishda yaxshiroq natija beradi. Har ikki model ham tilning semantik xususiyatlarini chuqur oʻrganishga moʻljallangan. Word2Vecning eng katta ustunligi shundaki, u soʻzlarning maʼnosini matematik vektorlarda aks ettira oladi. Masalan, erkak ayol juftliklari, davlat poytaxtlari, kasblar yoki fazoviy tushunchalardagi munosabatlarni geometrik tarzda ifodalash mumkin. Bunday imkoniyat Bag of Wordsda mavjud emas. Lekin Word2Vec oʻqitilishi ancha murakkab, katta korpus talab qiladi va hisoblash resurslari ham koʻproq talab etiladi.

Bag of Words va Word2Vecning taqqoslanishi

Har ikki model maqsadi matnni vektorga aylantirish boʻlsa ham, ularning natijalari va qoʻllanilish doirasi keskin farq qiladi. Bag of Words semantik maʼnoni tushunmaydi, Word2Vec esa aynan semantik aloqalarning oʻziga eʼtibor qaratadi. Bag of Words katta oʻlchamli, siyrak vektorlar yaratadi, Word2Vec esa kichik, zich va mazmunli vektorlar taqdim etadi. Bag of Wordsdan foydalanish oson, lekin murakkab matn tahlili talab qilinadigan joylarda yetarli boʻlmaydi. Word2Vec esa murakkab tizimlar uchun juda mos, masalan semantik qidiruv, kontekstual tavsiyalar, tarjima tizimlari, nutq tushunish, chatbotlar va intellektual muloqot tizimlarida keng qoʻllanadi.

Xulosa

Bag of Words matnni oddiy statistik model orqali ifodalasa, Word2Vec esa mazmunni chuqur o'rganadigan neyron tarmoqlar yordamida semantik vektorlar yaratadi. Kichik, sodda vazifalar uchun Bag of Words yetarli bo'lishi mumkin, ammo kontekst, ma'no va semantik bog'lanishni o'rganish zarur bo'lgan vazifalarda Word2Vec ancha samarali natija beradi. Shuning uchun zamonaviy NLP tizimlarida asosan Word2Vec yoki uning yanada rivojlangan modellari (FastText, GloVe, BERT) qo'llaniladi.

Foydalanilgan adabiyotlar

1. Mikolov, T., Chen, K., Corrado, G., & Dean, J. "Efficient Estimation of Word Representations in Vector Space." arXiv preprint arXiv:1301.3781, 2013.
2. Jurafsky, D., & Martin, J. H. "Speech and Language Processing." Pearson, 3rd Edition Draft, 2023.
3. Manning, C. D., Raghavan, P., & Schütze, H. "Introduction to Information Retrieval." Cambridge University Press, 2008.
4. Goldberg, Y. "Neural Network Methods for Natural Language Processing." Morgan & Claypool Publishers, 2017.
5. Harris, Z. "Distributional Structure." Word, 1954.
6. Rong, X. "Word2Vec Parameter Learning Explained." arXiv:1411.2738, 2014.