نام و نام خانوادگی:

کوییز ۳- درس پردازش زبان طبیعی

۱. همانطور که می‌دانید، یکی از مشکلات روش‌های Static Word Embeddings (یعنی آنهایی که contextual نیستند) درک درست از کلمات Polysemy است. ابتدا چند کلمه را مثال زده و توضیح دهید که چرا در این روش‌ها باعث چالش خواهند شد. سپس برای حل آن راه حل مناسبی را ارائه دهید. روش خود را به طور کامل توضیح دهید. (5 نمره)

پاسخ:

شیر(خوراکی - جنگل) - در این کلمات به ازای هر کلمه برداری ثابت در نظر گرفته می‌شود و همین عامل باعث آن شده که نتوانیم درک درستی از معنای کلمه داشته باشیم. استفاده از روش‌های Context-based مشکل فوق را مرتفع می‌کند.

۲. اولین مدل LM مبتنی بر شبکه عصبی که توسط Bengio, 2003 ارائه شد مبتنی بر feedforward است که در ان ۳ کلمه ورودی دارد و کلمه بعدی را پیش بینی میکند. در هر دو حالت که این شبکه از pretrained embedding استفاده بکند یا نکند، بررسی کنید که آیا این شبکه معادل یک 4gramاست؟ دلایل خود را با ذکر مثال بیاورید؟ (5 نمره)

پاسخ: در حالتیکه یک pre-trained embedding مانند w2v استفاده شود اکیدا پاسخ خیر است. شبکه خیلی قویتر است، بدلیل embedding که در شبکه استفاده میشود، این شبکه قوی تر از 4gramاست. اگر در ترین دیتا کلمه "cat" فقط داشته باشیم، انگاه در تست میتوان کلمه "dog" را بررسی کرد.در حالتیکه از prertained embedding استفاده نشود، نمیتوان با قطعیت قدرت ان را با 4gram مقایسه کرد

۳. فرض کنید مجموعه دادگانی در اختیار داریم که شامل چندین جمله است و می‌خواهیم آن‌ها را دسته بندی کنیم. برای آنکه بتوانیم دسته بندی را انجام دهیم نیاز است معیاری برای محاسبه شباهت محاسبه شود. با استفاده از بردارهای word2vec، معیار مناسبی را ارائه دهید. (5 نمره)

پاسخ:

میانگین گیری و جمع بردارهای آنها و سپس فاصله کسینوسی می‌تواند جزء راهکارهای مناسب باشد. از آنجایی که کلمات پر تکرار معنای خاصی به جمله وارد نمی‌کنند می‌توان آنها را یا نادیده گرفت و یا از اهمیت آنها کم کرد.

نام و نام خانوادگی:

کوییز ۳- درس پردازش زبان طبیعی

۴. با استفاده از ایده word2vec میخواهیم ایده ای node2vec را پیاده سازی کنیم. یعنی میخواهیم به گره های یک گراف یک بازنمایی برداری بسازیم بطوریکه گره های مشابه دارای بازنمایی های مشابه باشند. یک راه حل پیشنهاد دهید؟ هر گره دارای کلید یکتا است که ان را از بقیه گره ها متمایز میکند. گراف میتواند جهت دار یا بدون جهت باشد. (5 نمره)

پاسخ:

از روی گراف مذکور چندین مسیر بصورت تصادفی میسازیم. روش استانداردی برای تولید این مسیر با نام random walk وجود دارد. (اگر این اسم را نگفتید مهم نیست) . هر مسیر بصورت تصادفی از یک گره شروع میشود و به یک گره با طول رندم تمام میشود. هر مسیر نقش یک جمله را ایفاء میکند. کل مسیرهای تولید شده تشکیل یک کورپوس میدهند که توسط ان میتوان یک word2vec ترین کرد و درواقع هر گره دارای یک بازنمایی خواهد بود.

دو گره دارای بازنمایی یکسان است، اگر گره های همسایه یکسانی داشته باشند.