به نام خداوند تو

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

مبانی پردازش زبان و گفتار

تمرین دوم

استاد: دکتر حمیدرضا برداران کاشانی

دستیاران آموزشی:

مهرداد قصابی

علی مأمن پوش

محمدامین مولوی زاده

زمستان ۱۴۰۳ - بهار ۱۴۰۴

بخش اول: پرسش‌ها

۱- تفاوت بین one-hot-encoding و word embeddingها را بیان کنید

۲- به طور خلاصه توضیح دهید که الگوریتم GloVe چگونه Word Embeddingها را تولید می‌کند

۳- به طور خلاصه توضیح دهید که الگوریتم Word2Vec چگونه Word Embeddingها را تولید می‌کند.

۴- واژه‌های دارای چندمعنی چگونه در Word Embeddingها کنترل می‌شوند و چه چالش‌هایی را تولید می‌کند؟

۵- Word embeddingها نیاز دارند که تمام کلمات در مجموعه آموزش حضور داشته باشند. چگونه با OOV (Out of vocabulary) برخورد می‌کنید؟ یک روش برای تولید Word embedding برای کلماتی که درداده آموزش حاضر نبوده‌اند پیشنهاد دهید.

۶- ماتریس co-occurrence متن زیر را بنویسید. اندازه پنجره را ۲ در نظر بگیرید.

I love computer science and I love NLP even more.

۷- مزیت‌های نسبی GloVe و Word2Vec در چه کاربرد هایی بیشتر نمایان می‌شود؟

بخش دوم: Generative model using N-Gram

۱- با استفاده از پیکره Reuters مراحل زیر را انجام دهید

a. پیش‌پردازش
b. ساختن مدل با استفاده از N-Gram
c. تولید متن
d. ارزیابی (معیار Perplexity) و بهبودپارامترها
e. نتایج

بخش سوم: تحلیل احساسات با استفاده از Naïve Bayes

۱- مراحل زیر را با استفاده از movie_reviews انجام دهید:

۱-۱-پیش پردازش داده ها

۱-۲- ساختن مدل Naïve Bayes.

۱-۳- علایم نگارشی را حذف کنید.

۱-۴- ارزیابی مدل با استفاده از معیار های Accuracy, Precision, Recall, F1-score

بخش چهارم: شباهت معنایی

۱- مراحل زیر را با استفاده از پیکره Naab و پیاده‌سازی مقاله word2Vec انجام دهید:

a. پیش‌پردازش پیکره متنی (شامل توکن‌سازی، حذف Stop Words و...)
b. پیاده‌سازی مدل Word2Vec مبتنی بر Shallow Neural Network با یکی از روش‌های CBOW یا Skip-gram
c. آموزش مدل بر روی پیکره Naab و استخراج بردارهای کلمه. (۱۰ درصد پیکره برای آموزش کافی است)
d. تحلیل و تفسیر شباهت‌ها و ارائه مثال‌هایی از شباهت معنایی و تفاوت معنایی.
e. نمایش گرافیکی بردارهای کلمه در فضای دوبعدی (با استفاده از PCA یا TSNE)
f. مقایسه شباهت‌های حاصل با مدل های آماده مانند GloVe یاFastText)