پروژه درس پردازش زبان‌های طبیعی:

# خلاصه‌سازی متون با الگوریتم Sparse Coding

دکتر ممتازی
ترم بهار ۱۳۹۶-۱۳۹۷
دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر
زمان تحویل: ۲۵ تا ۲۷ تیر ۱۳۹۷

هدف از این پروژه استفاده از بازنمایی Term Frequency و Word2Vec برای پیاده‌سازی Summarization با کمک Sparse Coding می‌باشد.
فرض کنید که پیکره‌ای از تعدادی جمله داریم. در صورتی که هر جمله را به صورت یک بردار مانند Term-Frequency و یا میانگین Word2Vec کلمات آن جمله نشان دهیم، می‌توان پیکره را به صورت یک ماتریس که هر سطر آن بردار یک جمله است در نظر گرفت. به منظور به دست آوریدن خلاصه‌ای از این جملات می‌توان با کمک Sparse Coding یک مجموعه از سطرهای ماتریس را انتخاب کرد که با کمک آنها بتوان جملات دیگر را تقریب زد. به صورت دقیق‌تر اگر مجموعه جملات به دست آمده توسط Sparse Coding را $S^i=[s_1^i و s_2^i و \dots و s_k^i]$ بنامیم در آن صورت هر جمله $s_i$ (سطر مربوط به جمله $i$ ام در ماتریس اولیه) را می‌توان به صورت زیر تقریب زد.

$$s_i=\sum_{j=1}^{k} a_{ji} S_j^i$$

که در آنها مقادیر نامنفی هستند. در Sparse Coding به دنبال این هستیم که این تقریب کمترین خطا را داشته باشد.
برای جزئیات بیشتر می‌توانید به مقاله [1] مراجعه کنید.

در این پروژه قصد داریم از پیکره «پاسخ» استفاده کنیم. پیکره «پاسخ» (پیکره استاندارد سامانه های خلاصه ساز) در دو مدل تک سندی و چند سندی ارائه گردیده است. پیکره تک سندی شامل 100 موضوع مختلف از انواع گونه های خبری بوده که از خبرگزاری های پربیننده ایران انتخاب شده اند. هر کدام از این موضوعات دارای 5 خلاصه چکیده‌ای و استخراجی می باشند که توسط کارشناسان آموزش دیده تولید شده اند. پیکره چند سندی "پاسخ" نیز شامل 50 موضوع می‌باشد که هر موضوع حاوی 20 سند بوده و همچنین هر موضوع شامل 5 خلاصه انسانی و چکیده‌ای می باشد.
برای این پروژه از هر دو پیکره تک سندی و چند سندی استفاده کنید. یک بار خلاصه‌سازی بر روی پیکره تک سندی اجرا شود. در حالت تک سندی پیکره جملات ورودی را مجموعه جملات یک سند در نظر بگیرید. یک بار هم به‌صورت مجزا خلاصه‌سازی را بر روی پیکره چند سندی اجرا کنید. در حالت چند سندی پیکره جملات را مجموعه تمامی جملات در تمامی سندها در نظر بگیرید. برای کار با این پیکره

مرز جملات را نقطه، علامت سوال و علامت تعجب در نظر بگیرید. شایان ذکر است در برخی از خلاصه‌های گلد این پیکره مرز جملات هیچ یک از علامت‌های ذکر شده نیست و به صورت دستی مرز جملات مشخص شده اما برای پروژه نیاز نیست شما به علامتی غیر از علامت‌های گفته شده برای مرزبندی جملات توجه کنید.

مراحل انجام پروژه:

1. دیتاست پاسخ را از لینک زیر دانلود کنید:

https://goo.gl/hwiULH

2. برای داکیومنت‌های هر مجموعه، با استفاده از Term Frequency برداری تولید کنید.
3. الگوریتم Sparse Coding را بر روی بردارهای تولید شده اجرا نمایید.
4. خلاصه تولید شده، جمله‌های به دست آمده توسط Sparse Coding می‌باشد.
5. ارزیابی را با استفاده از معیار ROUGE1 و ROUGE2 بدست آورید و سپس بر اساس هر یک از دو معیار Recall، Precison و F-measure را محاسبه کنید.
6. با استفاده از روش Word2Vec برای واژگان موجود در دیتاست، یک بردار تولید کنید.
7. با توجه به بردارهای به دست آمده برای واژگان، بردار هر جمله را با میانگین‌گیری از مجموع بردارهای کلمات آن به دست آورید.
8. روش گفته شده برای Term Frequency را بر روی بردارهای جدید هم اجرا کرده و به طورمشابه ارزیابی نمایید.

منابع

[1] Multi-Document Summarization Based on Two-Level Sparse Representation Model