استخراج موضوع از شبکه‌های اجتماعی به کمک هشتگ‌ها

دانشجو: نوید رحیمی دانش

استاد راهنما: دکتر اسدپور

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران

مقدمه

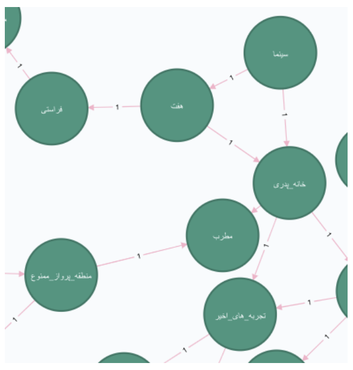
- ✓ در پردازش زبانی، استخراج موضوع، یا مدل موضوع، به روش‌های آماری برای یافتن موضوعات یا دسته‌بندی معنایی برای هر متن گفته می‌شود.
- ✓ در شبکه‌های اجتماعی، هشتگ‌ها برچسب‌های معنایی هستند که بوسیله کاراکتر (#) توسط کاربران مشخص می‌شوند.

در این تحقیق کاربردی، به هدف یافتن موضوعات مورد بحث در توییتر، یک روش موجود بهبود داده شد و به چندین حالت پیاده‌سازی گردید. در این پروژه موضوعات حول موضوع سینما مورد بررسی قرار گرفت. ورودی مساله مجموعه‌ای از توییت است و خروجی تعدادی موضوع مورد بحث در آن مجموعه توییت خواهد بود. روش پیاده سازی شده مبتنی بر گراف هشتگ‌ها است که بهتر از روش خام عمل می‌کند.

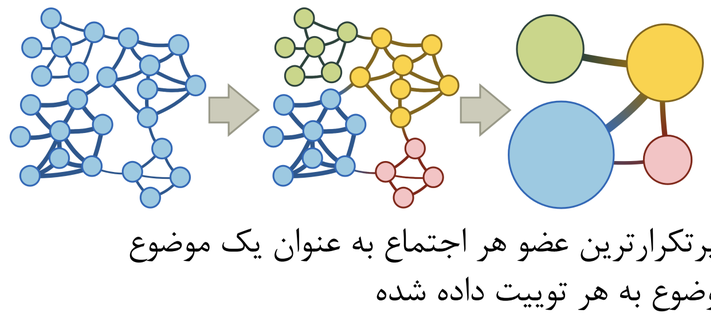
مدل پیشنهادی

در روش پیاده سازی شده مراحل زیر طی شدند:

1. جمع آوری تعداد متعددی توییت در یک بازه‌ی زمانی مشخص که حاوی کلماتی مانند (سینما، فیلم، ...) بودند
2. استخراج هشتگ‌ها و ساختن یک گراف وزن دار از آن‌ها که وزن یال نشان دهنده شباهت دو هشتگ خواهد بود



3. پیاده سازی دو معیار شباهت (f) برای دو هشتگ h_i و h_j
   - اگر دو هشتگ همزمان در یک توییت بیایند / در غیر اینصورت
$$f(h_i,h_j) = \begin{cases} 1, & \\ 0, & \end{cases}$$
   - استفاده از الگوریتم Labeled LDA که به کمک آن می‌توان بصورت احتمالاتی کلمات مرتبط با هر هشتگ را بدست آورد.
$$f(h_i, h_j) = -H(p(W_{h_i}), p(W_{h_j}))$$
   که $p(W_{h_i})$ همان توزیع احتمالی کلمات مرتبط با h_i است و H فاصله‌ی هلینگر آن دو توزیع است.
4. یافتن اجتماع‌های این گراف به کمک الگوریتم Louvain



5. انتخاب پرتکرارترین عضو هر اجتماع به عنوان یک موضوع
6. الصاق موضوع به هر توییت داده شده
7. ارزیابی:
   1. دقت موضوعات اختصاص داده شده به یک مجموعه توییت جدید
   2. میزان خلوص اجتماع‌های یافته شده به لحاظ تعداد هشتگ
   3. درصد موضوعات معنی‌دار یافته شده

❖ تمامی پیاده سازی ها به زبان پایتون انجام شده است.

نتایج

نتیجه اجرای مدل ما روی مجموعه‌ای از توییت‌های فارسی منتشر شده در بازه ۲۰ آبان تا ۱۵ آذر سال ۹۸ بصورت زیر بود:

- موضوعات تشخیص داده شده:
  - بنزین، سینما، سلبریتی_اجاره_ای، بازیگر، ایکسونامی، نوید_محمدزاده، زنان، تئاتر، مست_عشق، منطقه_پرواز_ممنوع، TheIrishman، اوج، تارانتینو، تحریم_سینما، IranProtests، رسانه، ماهشهر، صبا، نصرت_کریمی
- درصد موضوعات معنی‌دار تشخیص داده شده: ۶۳ درصد
- خلوص اجتماع‌های یافته شده: برای اکثر دسته‌ها خلوص بالا بود.
  - برای مثال گراف مقابل مربوط به موضوع TheIrishman است که تمامی هشتگ‌های موجود در آن به نحوی مرتبطند. یعنی خلوص ۱۰۰ درصدی داریم.

- روش داده از مرتبه زمانی n^2 است یعنی می‌تواند تا ده‌ها هزار هشتگ را در دقیقه پردازش کند.
- در مقایسه با روش LDA روش ما با استفاده از هشتگ‌ها موضوعات معنی‌دار بیشتری را پیدا خواهد کرد زیرا درصد خوبی از کاربران موضوع توییت را با هشتگ مشخص می‌کنند

جمع بندی

در این پروژه یک روش مبتنی بر گراف برای مساله استخراج موضوعات در شبکه‌های اجتماعی پیاده سازی شد. از روشی که در [1] معرفی شده شروع کردیم و با تغییر قسمت‌های معیار شباهت و الگوریتم شناسایی اجتماع به یک پیاده سازی قابل قبول رسیدیم. نتیجه را میتوان به راحتی در زبان پایتون اجرا کرد و موضوعات و دسته بندی هشتگ‌ها را در Neo4j مشاهده کرد.

کاربرد های صنعتی:

با توجه به گسترش استفاده از شبکه‌های اجتماعی شرکت‌ها و دولت‌ها علاقه‌مند به بررسی احساسات و نظرات حول موضوعات مختلف هستند که مساله استخراج موضوعات یک زیر مساله اساسی برای آن است.

همچنین برای دسته بندی مطالب می‌توان از این روش استفاده نمود.

مراجع اصلی

[1] Meng, X., Wei, F., Liu, X., Zhou, M., Li, S., & Wang, H. (2012, August). Entity-centric topic-oriented opinion summarization in twitter. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 379-387).

[2] De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2011, November). Generalized louvain method for community detection in large networks. In 2011 11th International Conference on Intelligent Systems Design and Applications (pp. 88-93). IEEE.