مدلسازی موضوعی از طریق الگوریتم LDA

با سلام و احترام

یه دیتاست متشکل از حدود ده هزار مقاله انگلیسی شامل عنوان، چکیده و سال انتشار از این مقالات دارم که لازمه عملیات زیر روی داده‌ها صورت بگیره:

پیش پردازش داده‌ها، استخراج کلمات کلیدی و تشکیل ماتریس واژه-سند بر اساس TF-IDF

خوشه‌بندی اسناد از طریق الگوریتم LDA

چند نمونه کار پیوست کردم، یکیشون پایان نامه است که من دقیقا کاری مشابه خروجی های این پایان نامه و دو تا مقاله نیاز دارم. صفحات پایانی این پایان‌نامه کدهای برنامه نویسی مربوط به اون پیوست شده که فکر میکنم به چگونگی پیاده سازی این الگوریتم کمک میکنه.

زمان و کیفیت کار برام خیلی مهمه. به کدهای برنامه نویسی و توضیحات و فیلم آموزشی روند انجام کار و چگونگی ارائه خروجی‌ها نیاز دارم تا بتونم خروجی هارو تفسیر کنم. به علاوه به کدهای برنامه نویسی هم نیاز دارم به این دلیل که باید پیوست پروژم بشن.

خروجی‌های مورد نیاز:

بخشی از ماتریس واژه- سند وزن‌دهی شده TF-IDF به عنوان نمونه؛

اطلاعات ماتریس شامل ابعاد ماتریس یعنی تعداد مدارک، تعداد کلیدواژه های استخراج شده، بیشترین طول لغات و کمترین طول لغات شامل چه تعداد کاراکتری است؟ معیار پراکندگی (measure of dispersion) چند درصد است؟ تعداد بلوک‌های خالی و تعداد بلوک‌های پُر؛

جدولی حاوی مجموعه لغات استخراج شده بر اساس TF-IDF همراه با بسامد واژگان؛

ترسیم ۱۰۰۰ تا از پرکاربردترین واژگان و واژگان هم رخداد یا نقشه ابری واژگان (Word Cloud) در دوره‌های زمانی مختلف. نتایج به صورت جدول هم باشه؛

تعیین خوشه‌های موضوعی به همراه ۱۵ تا از پرکاربردترین کلیدواژه‌های اختصاص یافته به هر خوشه در یک جدول.

معمولاً برای تعیین تعداد خوشه‌ها یا موضوعات از روش‌های زیر استفاده می‌شود:

۱. تعیین تعداد اولیه خوشه‌ها یا موضوعات و بعد با استفاده از روش نمونه‌گیری گیبس ( Gibbs Sampling method) دو دسته ده تایی قبل و بعد از آن تعداد در نظر گرفته شود

۲. مشخص کردن تعداد بهینهٔ موضوعات است بر اساس شاخص همبستگی

۳. تعیین تعداد خوشه‌ها با تعیین مقدار آلفا و دلتا

ارزیابی مدل موضوعی ایجاد شده با استفاده از معیارهای زیر:

گراف مربوط به انسجام مدلسازی[۱]؛ گراف احتمال[۲] (مشابهت و تناسب با دیگر موضوعات)؛ عدم یکدستی[۳] (نمونه کار مقاله‌ی "شناسایی موضوعات داغ و روندها در علم اطلاعات و دانش‌شناسی با استفاده از تکنیک‌های متن‌کاوی")

ترسیم نقشهٔ موضوعی یا نقشهٔ فاصله بین موضوعی (Intertopic distance map) برای نشان دادن مرتبط ترین گروه‌های کلمات در هر موضوع با استفاده از بسته نرم افزاری LDAVIS یا PyLDAvis (نمونه کار پایان نامه و مقالات پیوستی).

و نقشه ساختار مفهومی مطابق با نمونه کار (یک مقاله انگلیسی زبان که پیوست شده).

تعیین روند موضوعات داغ (موضوعاتی که در شرایط فعلی جز پژوهش‌های پرطرفدار هستند) و موضوعات سرد (موضوعاتی که زمانی جز موضوعات داغ بوده ولی در شرایط فعلی اقبالی به آن‌ها نیست و از میزان محبوبیت آن‌ها کاسته شده است) با استفاده از مدل رگرسیون خطی (liner regression model) و محاسبه پارامتر تتا و نمایش آن‌ها به صورت نموداری مطابق با نمونه کار

نتایج خوشه‌بندی موضوعی مقالات بر اساس سال انتشار آن‌ها در یک فایل اکسل.

[1] Topic coherence
[2] Likelihood
[3] Perplexity