بازیابی هوشمند اطلاعات

تمرین دوم

نام و نام خانوادگی : حسین سیفی

شماره دانشجویی : ۸۱۰۱۰۰۳۸۶

آذر ۱۴۰۰

فهرست

۱-هدف تمرین

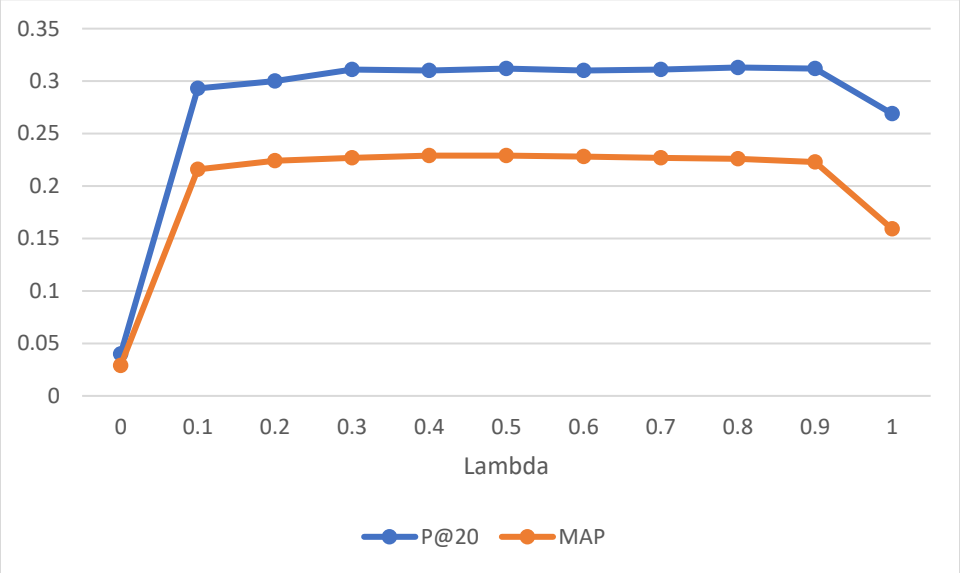
- مطالعه روش‌های مختلف هموارسازی توابع بازیابی و پارامترهای آن
- بسط پرسوجوی کاربر با استفاده از روشPseudo Relevance Feedback

۲- سوال ۱

یکی از مشکلات مطرح در حوزه بازیابی اطلاعات، وجود احتمالهای صفر است که محاسبات را درعمل دچار مشکل میکند. روشهای هموارسازی برای حل این مشکل مطرح شدند تا احتمال رخداد کلمات دیده نشده پرسوجو در اسناد را تخمین بزنند. در این سوال روش JM ، Dirichlet-prior و additive-smoothing مورد بررسی قرار می‌گیرند.

۲٫۱- روش JM

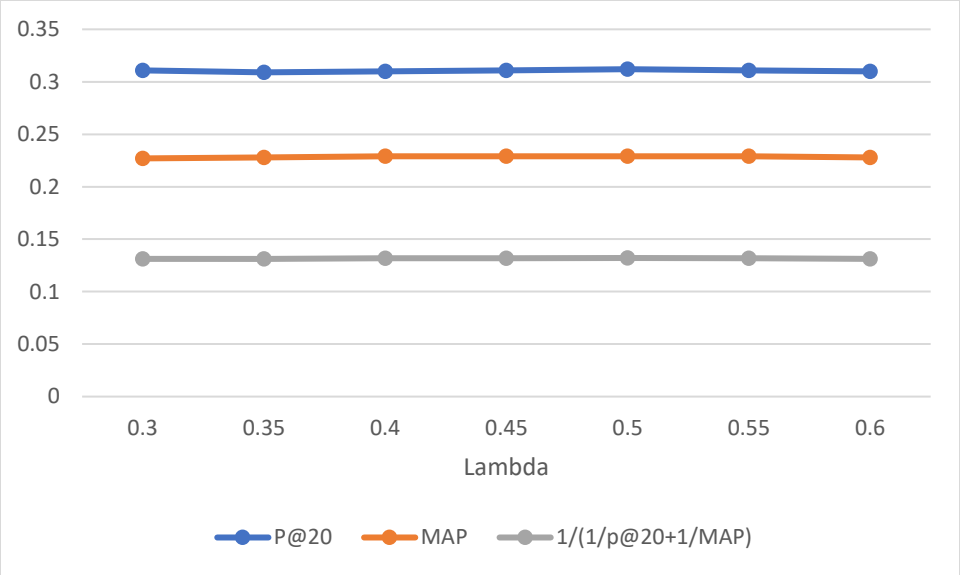
در ابتدا گام‌های بلندی(۰.۱ واحدی) برای lambda که با توجه به فرمول باید مقداری بین ۰ و ۱ داشته باشد در نظر می‌گیریم و نتایج معیارهای ارزیابی را به ازای مقادیر متفاوت این پارامتر به دست می‌آوریم. سپس محدوده‌ای از مقادیر این پارامتر را مشخص می‌کنیم که به ازای آن مقادیر معیارهای ارزیابی بهترین نتایج را نشان می‌دهند. نمودار معیارهای ارزیابی MAP و P@20 برای مقادیر Lambda گفته شده به شکل زیر است.



نمودار ۱

با توجه به نمودار ۱ مشاهده می‌شود که بهترین مقادیر هر دو معیار ارزیابی MAP و P@20 بین lambda=0.3 و lambda=0.6 قرار دارد در نتیجه این بار بازه انتخاب شده با گام‌هایی کوتاه‌تر(۰.۰۵ واحدی) طی می‌شود تا بهترین مقدار lambda برسیم. نمودار زیر مقدار معیارهای ارزیابی در بازه انتخاب شده نشان می‌دهد.

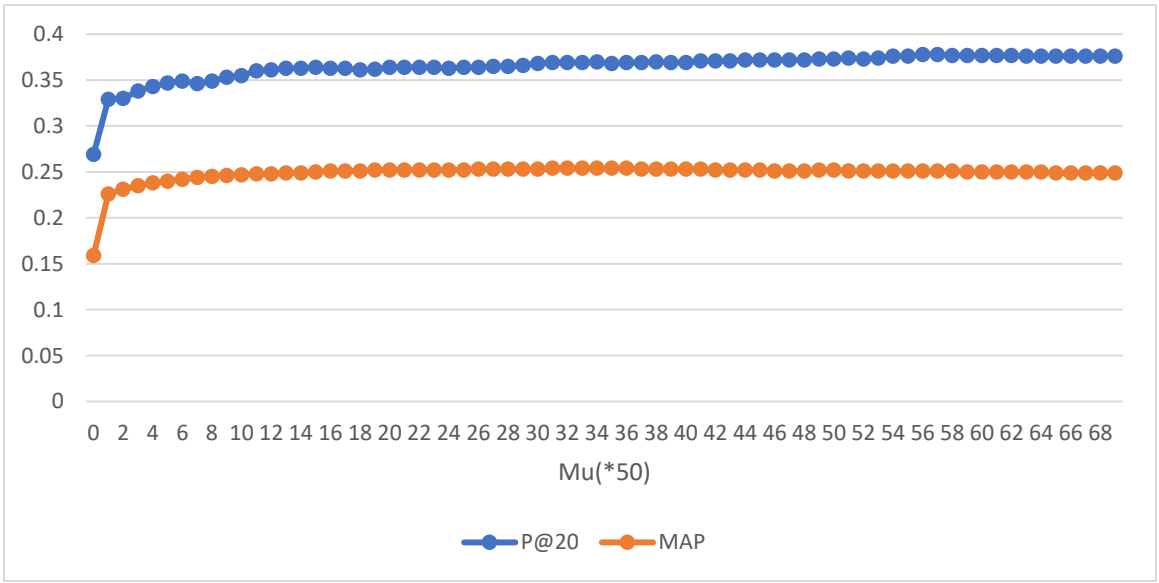
با دقت در نمودار ۲ و استفاده از ترکیب مقادیر ارزیابی به وسیله فرمول 1/(1/p@20+1/MAP) می‌توان متوجه شد که در نقطه lambda=۰.۵ بهترین مقادیر معیارهای ارزیابی را داریم.



نمودار ۲

۲٫۲- روش Dirichlet-Prior

پارامتر Mu موجود در فرمول روش Dirichlet-Prior باید عددی حقیقی نامنفی باشد در نتیجه در ابتدا باید گام‌های بسیار بلندی (۵۰ واحد) از ۰ تا سقف ۳۵۰۰ بررسی می‌شود. نمودار ۳ ، معیارهای ارزیابی MAP و P@20 برای روش Dirichlet-Prior به ازای مقادیر متفاوت پارامتر Mu را نشان می‌دهد.

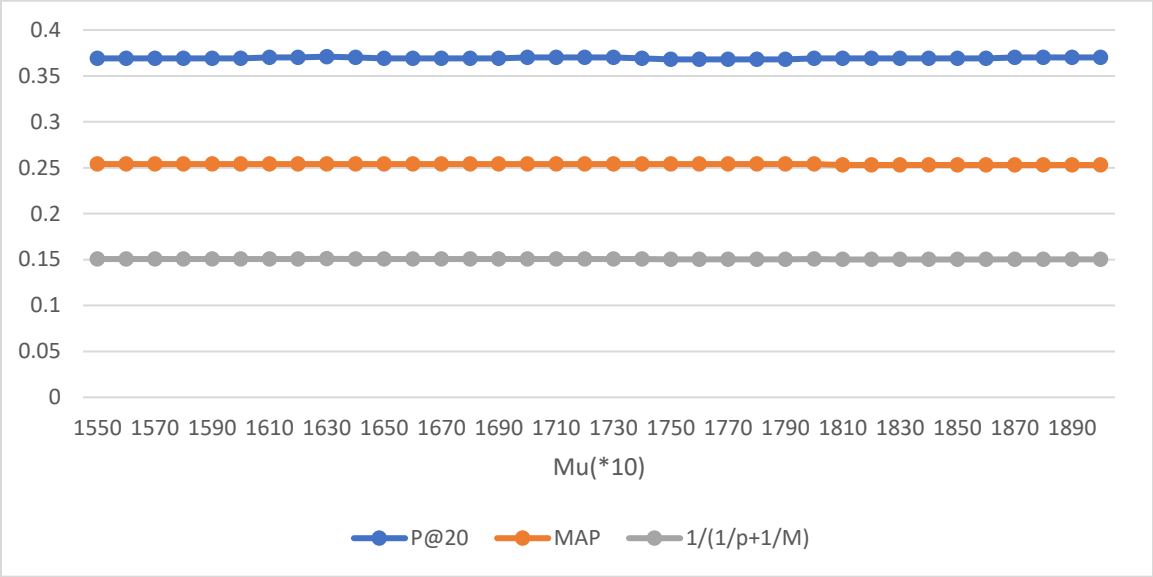


نمودار ۳

با توجه به نمودار ۳ ، مشاهده می‌شود در صورتی که مقدار Mu بین ۱۵۵۰ و ۱۹۰۰ باشد بهترین مقادیر معیار ارزیابی MAP به دست می‌آید و در صورتی که مقدار Mu بین ۲۷۵۰ و ۲۹۰۰ باشد ، بهترین مقادیر معیار ارزیابی P@20 به دست می‌آید. همانطور که توصیف شد ، این دو بازه [۱۵۵۰,۱۹۰۰] و [۲۷۵۰,۲۹۰۰] هیچ اشتراکی با یکدیگر ندارد و ممکن است استفاده از فرمولی که

برای ترکیب معیارهای ارزیابی در سوال قبل استفاده شد برای این سوال معقول نباشد و از آنجایی که معیار ارزیابی MAP معیار مهمتری است ، با اعمال فرمولِ ترکیب ، نقاطی به دست می‌آید که مقدار ایده‌آلی از MAP را ارائه نمی‌دهند در نتیجه بهتر است که برای به دست آوردن مقدار دقیقتر Mu روی بازه انتخابی با توجه به MAP تمرکز کنیم.

در نمودار ۴ تغییرات معیارهای ارزیابی نسبت به Mu بین مقادیر ۱۵۵۰ و ۱۹۰۰ با گام‌های کوتاه(۱۰ واحد) نشان داده شده است.



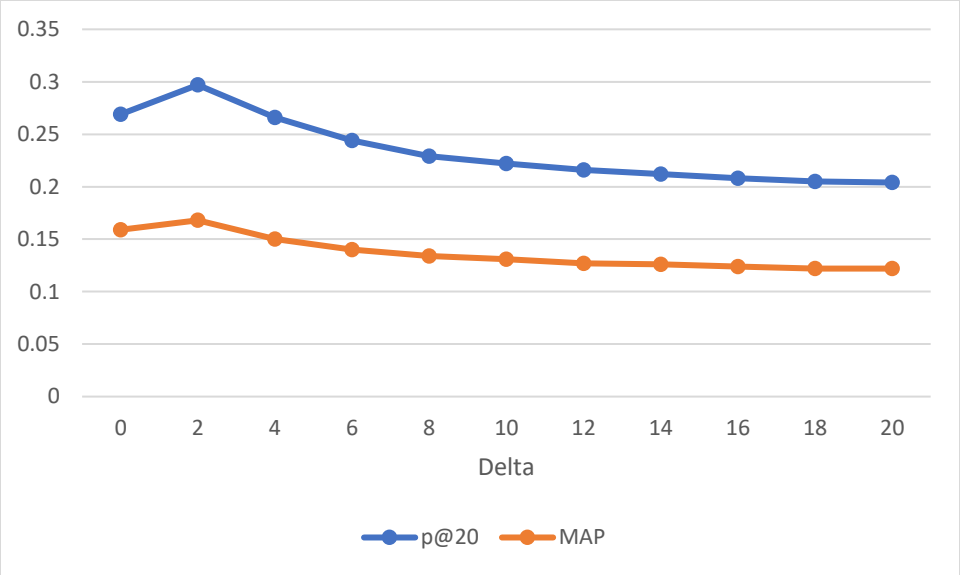
نمودار ٤

با توجه به نمودار ۴ ، تغییرات Mu در این بازه باعث تغییر چندانی در معیارهای ارزیابی نمی‌شود اما با بررسی بیشتر و به کمک فرمول 1/(1/p@20+1/MAP) می‌توان مشاهده کرد که در نقطه Mu=۱۶۳۰ بهترین مقدار از ترکیب معیارهای ارزیابی به دست می‌آید.

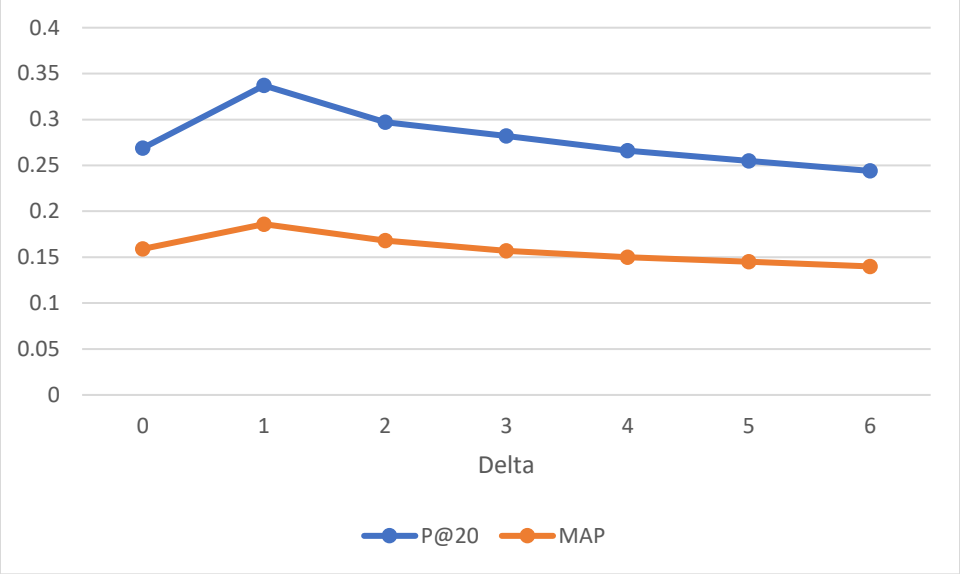
۲٫۳- روش Additive-smoothing

این روش به صورت پیش‌فرض در Galago موجود نبود و با تغییر توابع موجود در فایل DirichletScoringIterator.java روش Additive-Smoothing را ایجاد کردیم. در این فرمول پارامتر Delta وجود دارد که باید عددی صحیح و نامنفی باشد چرا که از نظر شهودی ، دلتا تعدادی است که از ترم‌های موجود در سند کم می‌کنیم تا به ترم‌هایی که در سند موجود نیستند اختصاص دهیم. در نمودار زیر ، دلتا در بازه ۰ تا ۲۰ با گام‌هایی نسبتا بلند(۲ واحدی) می‌شود.

در نمودار ۵ ، هر دو معیار MAP و P@20 در نقطه delta=۲ بهترین مقدار را دارند اما با بررسی دقیق‌تر ، با گام‌هایی کوتاه(۱ واحد) در بازه delta=۰ تا delta=۶ می‌توان بهترین مقدار دلتا برای مطلوب‌ترین نتیجه را یافت. نمودار ۶ بررسی معیارهای ارزیابی در بازه گفته شده را نشان می‌دهد.



نمودار ۵



نمودار ۶

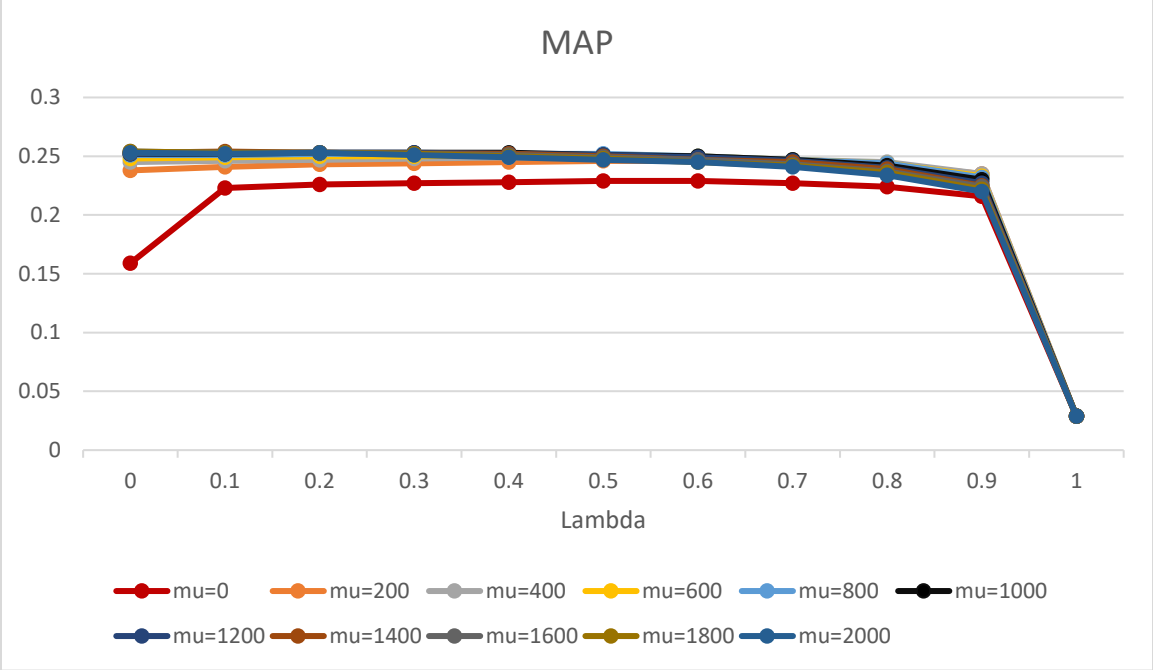
نمودار ۶ به خوبی نشان می‌دهد که به درستی به نتیجه بررسی بازه دلتا با گام‌های بلند اکتفا نکردیم و بازه‌ی مشکوک را با گام‌هایی نسبتا کوتاه‌تر بررسی کردیم. در این نمودار مشخص است که هر دو معیار ارزیابی MAP و P@20 بهترین مقدار را برای پارامتر دلتا عدد ۱ می‌دانند. به عبارتی دیگر ، روش هموارسازی لاپلاس از هموارسازی Additive با دلتاهایی مخالف ۱ بهتر عمل می‌کند.

۳- سوال ۲

در این سوال ابتدا پیاده سازی بازیابی به کمک روش هموارسازی دو مرحله‌ای(Two-Step-Smoothing) انجام شده است که طبق خواسته سوال ، کد توضیح داده نمی‌شود.

می‌دانیم که این روش شامل دو پارامتر Mu و Lambda می‌باشد ، پس نیاز داریم با روشی مشابه سوال قبل بهترین مقادیر را برای این پارامترها انتخاب کنیم تا معیارهای ارزیابی نتایج مطلوب‌تری را نسبت به دیگر مقادیر پارامترها به ما نشان دهند.

در ابتدا به وسیله‌ی حلقه‌ای تودرتو و بازه ۰ تا ۲۰۰۰ برای Mu با گام‌هایی برابر ۲۰۰ و بازه ۰ تا ۱ برای Lambda با گام‌هایی ۰.۱ واحدی معیارهای ارزیابی را به دست می‌آوریم . دلیل استفاده از حلقه تودرتو این است که هر مقدار از Mu با مقادیر متفاوتی از Lambda ارزیابی شود و بالعکس.



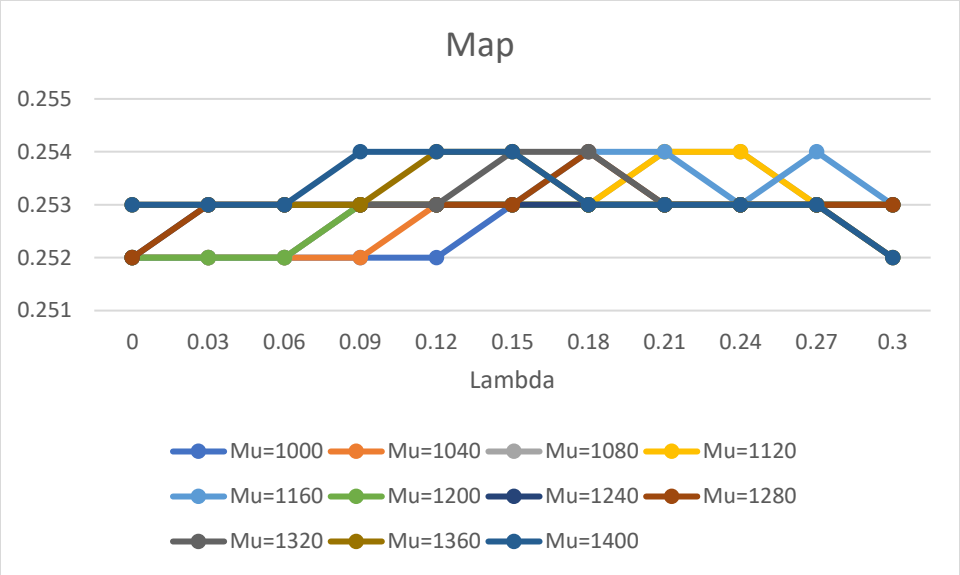
نمودار ۷

بررسی تمامی مقادیر معیارهای ارزیابی MAP و P@20 برای این سوال می‌تواند بسیار پیچیده باشد در نتیجه برای اندکی ساده سازی فقط معیار MAP را بررسی می‌کنیم که مقادیر آن با توجه به تغییرات Mu و Lambda در نمودار ۷ نشان داده شده است.

در نمودار ۷ هر ۱۱ مقدار انتخاب شده برای Mu شباهتی در نقطه lambda=۱ دارند. دلیل آن هم این است که با انتخاب مقدار ۱ برای پارامتر lambda ، بازیابی کاملا از سند مورد بررسی مستقل شده و تنها به Language Model وابسته می‌شود. در سایر نقاط نیز قابل مشاهده است که به ازای lambda=۰ تا lambda=۰.۳ وهمچنین Mu=۱۰۰۰ تا Mu=۱۴۰۰ بهترین مقادیر معیار ارزیابی MAP به وجود می‌آید در نتیجه در مرحله بعد با گام‌های ۰.۰۳ واحدی برای lambda و گام‌های ۴۰ واحدی برای Mu بازه‌های مشخص شده را می‌پیماییم تا بهترین نتایج را به دست آوریم.

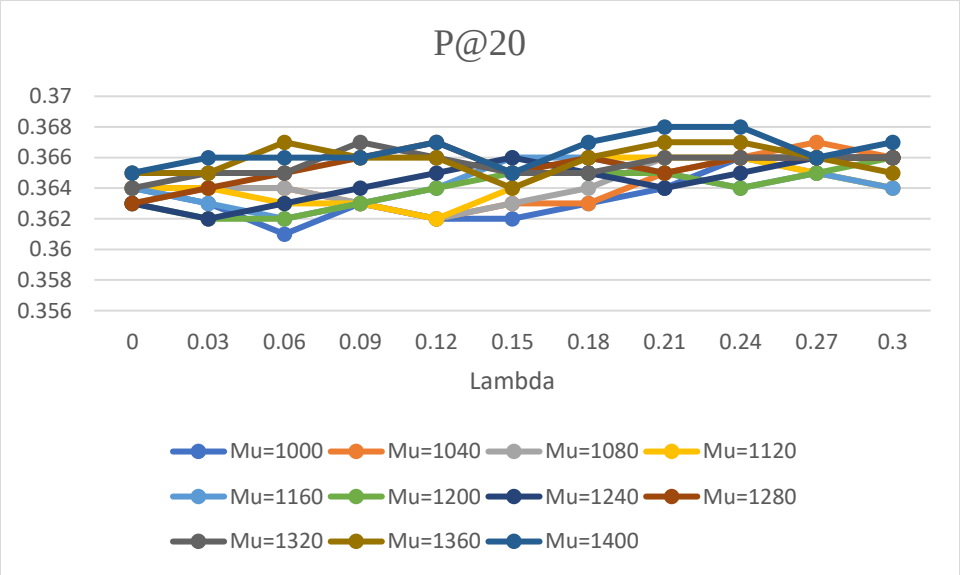
مقادیر معیارهای ارزیابی MAP برای بازه‌های ذکر شده در نمودار ۸ مشخص است. در این نمودار تعداد زیادی از نقاط دارای مقداری برابر و حداکثری دارند. از این گزاره می‌توان دو احتمال را مطرح کرد:

۱. نرخ نمونه برداری(Sampling Rate) زیاد است.
۲. MAP به تنهایی کافی نیست و باید از معیارهای ارزیابی دیگری(مانند P@20) استفاده کنیم.



نمودار ۸

در نتیجه برای تصمیم گیری و انتخاب بهترین مقادیر پارامترها ، نمودار ۹ که معیار ارزیابی P@20 را برای بازه‌های مشابهی با نمودار ۸ نشان می‌دهد ، رسم می‌کنیم.



نمودار ۹

با در نظر گرفتن هر دو نمودار ۸ و نمودار ۹ درمی‌یابیم که تنها نقطه‌ای که مقدار MAP حداکثری(۰.۲۵۴) دارد و مقدار P@20 متفاوتی با بقیه نقاط دارد(۰.۳۶۷) دارای پارامترهای lambda=۰.۱۲ و Mu=۱۴۰۰ است.