# پروژه درس بازیابی اطلاعات

امير حسين سرور - ٩٧٣١٠٢٨

# گزارش فاز سوم

۱) سه پرسمان مناسب را انتخاب کرده، نتایج را از نظر عملکرد و سرعت موتور جستجو برای این سه پرسمان در دو حالت
بدون خوشه بندی و با خوشه بندی مقایسه و تحلیل نمایید.

پرسمان: سازمان ملل متحد

نتايج:

# با خوشه بندی (b=5)

### بدون خوشه بندى

```
>> Please Enter your Query: ﴿ 8 30 24 X المعدد المعدد المعدد المعدد حدد المعدد المعدد
```



# پرسمان: انتخابات ریاست جمهوری 🕨

## با خوشه بندی (b=10)

```
>> Please Enter your Query: المراد مال المراد المراد المرد على المرد ال
```

#### بدون خوشه بندي

# 🕨 پرسمان: تفریغ بودجه



## با خوشه بندی (b=1)

#### بدون خوشه بندي

```
>> Flease Enter your Query: الله عَهِلَّهُ عَهِلَهُ اللهُ عَهِدِهُ اللهُ عَهْدِهُ اللهُ عَهْدُهُ اللهُ عَمْدُهُ اللهُ اللهُ عَمْدُهُ اللهُ اللهُ عَمْدُهُ اللهُ عَمْدُهُ اللهُ عَمْدُهُ اللهُ اللهُولِ اللهُ ا
```

# تحلیل، مقایسه و توضیحات:

در حالت خوشهبندی، جستجو با طهای مختلف (یعنی هر کوئری با اسناد چندخوشه مرتبط مقایسه شود) را امتحان کردیم و نکته واضح آن است با افزایش فی را بازیابی افزایش پیدا می کند چرا که جستجو با خوشهها و به طبع با اسناد بیشتری مقایسه می شود و همچنین اینکه زمان بازیابی در حالت خوشهبندی شده نسبت به حالت بدون خوشهبندی، از آنجایی که جستجو با اسناد کمتری صورت می گیرد و لازم به پیمایش تمام اسناد نیست، به طور کلی زمان کمتری دارد.



از لحاظ کیفیت نتایج، خوشهبندی اسناد عملکرد رضایت بخشی به همراه داشته است. با توجه به تعداد زیاد اسناد (۵۰ هزار) و اینکه در الگوریتم kmean تعداد خوشهها را در  $\gamma$  حالت  $\gamma$  حالت  $\gamma$  الله الله الله و الله الله و می تواند داشته باشد که ما حتی در صورتی که  $\gamma$  الله و از در خود می تواند داشته باشد که ما حتی در صورتی که  $\gamma$  اسناد بسیار مرتبطی را مشاهده می کنیم. کوئری با اسناد یک خوشه (شبیه ترین خوشه)، در نتایج کوئری «تفریغ بودجه» اسناد بسیار مرتبطی را مشاهده می کنیم. با توجه به اینکه در حالت بدون خوشه بندی، کوئری با تمامی اسناد مقایسه شده و نتایج بر اساس میزان شباهت مرتب می شوند و مطمئن هستیم اسنادی که در نتایج بالاتری نیستند، شباهت بیشتری هم نداشته اند، و در حالتی که خوشه بندی می کنیم کوئری با مجموعه زیادی از اسناد دیگر مقایسه نخواهد شد، ولی می بینیم که نتایج هردو رضایت بخش هستند و خوشه بندی از این نظر که نوعی هرس روی تمام اسناد انجام داده و دیگر اسناد نامر تبط به مرحله مقایسه با کوئری راه خوشه بندی از این وجود عملکرد خوبی هم خواهیم داشت، الگوریتم کار آمدی محسوب می شود.

۲) به ازای هر دسته یک پرسمان مناسب انتخاب کنید و نتایج جستجوی این پرسمان را در دو حالت با دستهبندی و بدون
دستهبندی مقایسه و تحلیل کنید.

- ❖ دسته ورزشی (sport):
  - **پرسمان:** دادگاه



در حالت بدون دستهبندی، با توجه به پرسمان «دادگاه»، نتایجی که بازیابی شده با اینکه مرتبطند، اما از دستههای مختلفی مثل سیاسی، فرهنگی و ورزشی تشکیل شدهاند. در حالتی که دستهبندی داریم و پرسمان «دادگاه» فقط در اسناد مربوط به دسته ورزشی جستجو می شود، نتایجی که بازیابی شدهاند کاملا مربوط به وقایع ورزشی و همچنین وقایعی که در آنها «دادگاه» نقش پررنگی دارد، مثلا وقایع قضایی در ورزش، هستند.

- دسته اقتصادی (economy):
  - 🗸 پرسمان: ميوه



در حالت اول، نتایجی که برای پرسمان «میوه» بازیابی شدهاند، از دسته های سلامت، فرهنگی و اقتصادی میباشند. قابل توجه است خبر اقتصادی مربوط به «میوه» آخرین سندی است که در حالت بدون دسته بندی بازیابی شده است. در حالتی که با استفاده از دسته بندی و جستجوی «میوه» در دسته اقتصادی، همان خبر به عنوان اولین خبر بازیابی شده و بقیه اخبار هم اسنادی مرتبط با پرسمان و همچنین دارای مفهومی اقتصادی هستند.

- 🌣 دسته سیاسی (political):
  - 🖊 پرسمان: استقلال



```
>> Please Enter your Query: ماتهادات المات الما
```

در نگاه اول، شاید کلمه «استقلال» ما را به یاد اسناد ورزشی بیندازد، که این مورد را در حالت بدون دستهبندی نیز مشاهده می کنیم. از آنجایی که پرسمان «استقلال» در اخبار ورزشی تکرار بیشتری داشته است، تمرکز اسناد بازیابی شده بر روی اخبار ورزشی و مرتبط با استقلال است. اما اگر همین پرسمان را با استفاده از دستهبندی و در دسته سیاسی جستجو کنیم، نتایجی که بازگردانده می شوند تماماً مربوط به مفاهیم سیاسی «استقلال» هستند که در این مورد را می توانیم در نتایج بازیابی شده با دستهبندی مشاهده کنیم.

- 💠 دسته فرهنگی (culture):
  - 🗸 پرسمان: اوراسیا



پرسمان «اوراسیا» در حالت عادی، هم پرسمانی نسبتاً نادر محسوب میشود هم اینکه اسناد بازیابی شده برای آن، از دستههای مختلف فرهنگی، سیاسی، اقتصادی و حتی ورزشی میتواند باشد که در حالت بدون دستهبندی این به همین صورت است. در حالت با دستهبندی، اسنادی بازیابی شدهاند که مربوط به وقایع فرهنگی پرسمان «اوراسیا» میشوند که در تصویر بالا قابل مشاهده است.

- دسته سلامت (health):
  - 🔻 **پرسمان:** افطار



در حالت بدون دستهبندی، اسنادی که برای پرسمان «افطار» بازیابی شدهاند، از دستههای مختلف سلامت، فرهنگی، ورزشی و سیاسی هستند، از آنجهت که این پرسمان به موارد متعددی در وقایع مربوط به ماهرمضان می تواند اشاره داشته باشد. اما اگر بخواهیم این جستجو را محدود کنیم و به جنبه سلامت «افطار» بپردازیم، در حالت دوم که از دستهبندی استفاده شده می توان دید که اسناد بازیابی شده تماماً مربوط به جنبه سلامت این پرسمان هستند، مثلاً توصیههای غذایی و درمانی مربوط به روزه، ماه رمضان و «افطار» در میان اسناد بازیابی شده مشاهده می شود که قابل انتظار است.

