بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

---

**بازیابی پیشرفته‌ی اطلاعات**
نیم‌سال دوم تحصیلی ۰۱-۰۰

# تمرین سوم: سامانه‌ی جستجو نوشته‌های مرتبط با سلامت/بیو

محمدمهدی ابوترابی ۹۸۱۰۵۵۵۷
یاسمن زلفی موصلو ۹۸۱۰۵۷۹۵
فاطمه عسگری ۹۸۱۰۵۹۲۱

## مقدمه

در این تمرین ابتدا داده‌های مرتبط به سلامت از تارنماهای اخبار و نوشته‌های سلامت جمع‌آوری شده و پس از انجام پیش‌پردازش‌های لازم بر روی داده، سامانه‌ی جستجویی برای مطالب پزشکی و مرتبط با سلامت طراحی شده است. این سامانه با استفاده از چهار روش بازیابی boolean، روش tf-idf، روش مبتنی بر transformerها و همچنین روش fasttext، با دریافت یک موضوع یا عنوان از کاربر، تعدادی نوشته‌ی مرتبط را به او پیشنهاد می‌دهد.

## ساختار کلی پروژه

در این بخش به توضیح کلی قسمت‌های مختلف پروژه و کاری که هر یک انجام می‌دهند می‌پردازیم:

**پوشه‌ی crawlers:** این پوشه حاوی فایل‌هایی است که در آن‌ها داده‌های مرتبط با سلامت در تارنماهای **نمناک** و **سلام دکتر** را دریافت کرده‌ایم. برای افزایش داده‌ها، علاوه بر داده‌هایی که به صورت آماده به ما داده شده بود، به همان اندازه داده کراول کردیم و در کنار داده‌های داده‌شده قرار دادیم.

**پوشه‌ی dataset:** داده‌های مورد نیاز پس از جمع‌آوری و کراول کردن، بر حسب نیاز تمیز شده و اطلاعات اضافی آن مانند تگ‌های موجود حذف شده و در قالب فایل‌های json داخل این پوشه قرار داده شد. تمامی مدل‌های استفاده شده به جز مدل transformer تنها از این مجموعه داده استفاده می‌کنند.

**پوشه‌ی old-dataset:** این پوشه شامل داده‌های جمع‌آوری شده قبل از تمیز کردن است. لازم به ذکر است که مدل transformer علاوه بر dataset تمیز از این دیتاست نیز استفاده می‌کند.

**پوشه‌ی notebooks:** این پوشه حاوی چهار فایل نوت‌بوک است که در هر کدام یکی از مدل‌ها و روش‌های استفاده‌شده پیاده‌سازی شده است؛ یعنی چهار روش boolean، tf-idf، fasttext، و trsnsformer

۱- روش boolean

در این روش ابتدا بر روی داده‌ی موجود پیش‌پردازش‌های لازم از جمله نرمالایز کردن، توکنایز کردن، حذف stop wordها، و همچنین lemmatize کردن انجام شده است. سپس ماتریس term-docای که ستون‌هایش مربوط به کلمات و سطرهایش مربوط به سندها هستند ساخته و در یک فایل ذخیره شده تا در ادامه مدل آن را لود کرده و از آن استفاده کند.

کوئری‌هایی که به این مدل داده می‌شود در قالب کلماتی است که بین آن‌ها عملگر and یا or آمده است. همچنین ممکن است پیش از یک کلمه not آمده باشد که به معنی نیامدن آن کلمه در سند بازگردانده شده است. در این مدل ابتدا عملگرها و عملوندها مشخص شده و برای هر عملوند، بردار مرتبط با آن از درون ماتریس بولین برداشته شده و با بردار مربوط به عملوند بعدی، با توجه به عملگر بین آن دو، and یا or می‌شود. در نهایت سطرهایی از بردار که مقدار true دارند شناسایی شده و سند متناظر با آن‌ها به عنوان سند مرتبط بازگردانده می‌شود.

یکی از نتایج این روش در زیر مشاهده می‌شود. نتایج کامل در نوت‌بوک مربوطه موجود است:

```
BooleanModel().print_similars('کرونا and ریه')
```
Python

```
Output exceeds the size limit. Open the full output data in a text editor
1- title: آنچه باید درباره عفونت همزمان کرونا و آنفلوانزا بدانید
1- link: https://www.hidoctor.ir/357168_%d8%a2%d9%86%da%86%d9%87-%d8%a8%d8%a7%db%8c%d8%af-%d8%af%d8%b1%d8%a8%d8%a7%d8%b1%d9%87-%d8%b9%d9%81%d9%88%d9%86%d8%aa-%d9%87%d9%85%d8%b2%d9%85%d8%a7%d9%86-%da%a9%d8%b1%d9%88%d9%86%d8%a7-%d9%88-%d8%a2.html/
--------------------------
2- title: در مورد بیماری کرونا ویروس چه می‌دانید؟
2- link: https://www.hidoctor.ir/346899_%d8%af%d8%b1-%d9%85%d9%88%d8%b1%d8%af-%d8%a8%db%8c%d9%85%d8%a7%d8%b1%db%8c-%da%a9%d8%b1%d9%88%d9%86%d8%a7-%d9%88%db%8c%d8%b1%d9%88%d8%b3-%da%86%d9%87-%d9%85%db%8c-%d8%af%d8%a7%d9%86%db%8c%d8%af%d8%9f.html/
--------------------------
3- title: علائم و عوارض بلند مدت کرونا + حقایق ناگفته
3- link: https://namnak.com/coronavirus-long-term-effects.p83159
--------------------------
4- title: علائم نشان دهنده انتشار ویروس کرونا در ریه‌ها
4- link: https://www.hidoctor.ir/352292_%d8%b9%d9%84%d8%a7%d8%a6%d9%85-%d9%86%d8%b4%d8%a7%d9%86-%d8%af%d9%87%d9%86%d8%af%d9%87-%d8%a7%d9%86%d8%aa%d8%b4%d8%a7%d8%b1-%d9%88%db%8c%d8%b1%d9%88%d8%b3-%da%a9%d8%b1%d9%88%d9%86%d8%a7-%d8%af%d8%b1.html/
```

۲- روش tf-idf:

در این روش نیز ابتدا بر روی داده‌ی موجود پیش‌پردازش‌های لازم از جمله نرمالایز کردن، توکنایز کردن، حذف stop wordها، و همچنین lemmatize کردن انجام شده است. سپس این روش از tfIdfVectorize استفاده کرده است که یک ماتریس می‌سازد و برای هر سند یک بردار به دست می‌آید. این ماتریس نیز ابتدا ذخیره و در ادامه برای استفاده لود شده است. همچنین این روش دارای یک مدل است که آن هم داخل پوشه‌ی مربوطه ذخیره شده است.
مدل پیاده‌سازی‌شده، وکتور کوئری داده‌شده را در همان فضای برداری می‌سازد و سپس فاصله‌ی کسینوسی آن را با تمام وکتورهای سندها محاسبه کرده و k نزدیکترین را برمی‌گرداند.
یکی از نتایج این روش در زیر مشاهده می‌شود. نتایج کامل در نوتبوک مربوطه موجود است:

```
TfidfModel().print_similars('ویروس کرونا')
```
Python

```
Output exceeds the size limit. Open the full output data in a text editor
1- title: علائم ویروس کرونا در کودکان چه تفاوتی دارد؟
1- link: https://www.hidoctor.ir/352565_%d8%b9%d9%84%d8%a7%d8%a6%d9%85-%d9%88%db%8c%d8%b1%d9%88%d8%b3-%da%a9%d8%b1%d9%88%d9%86%d8%a7-%d8%af%d8%b1-%da%a9%d9%88%d8%af%da%a9%d8%a7%d9%86-%da%86%d9%87-%d8%aa%d9%81%d8%a7%d9%88%d8%aa%db%8c-%d8%af.html/
--------------------------
2- title: در مورد بیماری کرونا ویروس چه می‌دانید؟
2- link: https://www.hidoctor.ir/346899_%d8%af%d8%b1-%d9%85%d9%88%d8%b1%d8%af-%d8%a8%db%8c%d9%85%d8%a7%d8%b1%db%8c-%da%a9%d8%b1%d9%88%d9%86%d8%a7-%d9%88%db%8c%d8%b1%d9%88%d8%b3-%da%86%d9%87-%d9%85%db%8c-%d8%af%d8%a7%d9%86%db%8c%d8%af%d8%9f.html/
--------------------------
3- title: چطور بفهمم دلیل سردردم کروناست؟
3- link: https://namnak.com/headache-coronavirus-symptom.p81893
--------------------------
4- title: شایع‌ترین علائم پس از ابتلا به کرونا / چه زمانی باید درخواست کمک کنیم؟
4- link: https://www.hidoctor.ir/354939_%d8%b4%d8%a7%db%8c%d8%b9-%d8%aa%d8%b1%db%8c%d9%86-%d8%b9%d9%84%d8%a7%d8%a6%d9%85-%d9%be%d8%b3-%d8%a7%d8%b2-%d8%a7%d8%a8%d8%aa%d9%84%d8%a7-%d8%a8%d9%87-%da%a9%d8%b1%d9%88%d9%86%d8%a7-%da%86%d9%87.html/
```

۳- روش fastext:

در این روش نیز ابتدا بر روی داده‌ی موجود پیش‌پردازش‌های لازم از جمله نرمالایز کردن، توکنایز کردن، حذف stop wordها انجام شده است. در این روش مدل fasttext ساخته شده و متن‌های توکنایزشده‌ی سندها به این مدل داده می‌شود. با میانگین گرفتن از روی وکتورهای کلمات متن، وکتور مربوط به هر سند ساخته می‌شود. (این مدل به وکتور متن وزن ۹ و به وکتور عنوان وزن ۱ داده و میانگین وزن‌دار می‌گیرد و وکتور حاصل را به عنوان وکتور مربوط به سند نگه می‌دارد). سپس مدل ذخیره می‌شود تا در ادامه لود شده و استفاده گردد.
در ادامه وقتی کوئری داده می‌شود، به هر کوئری نیز یک وکتور نسبت می‌دهد. در نهایت مانند روش tf-idf فاصله‌ی کسینوسی بردار کوئری با هر سند محاسبه شده و k نزدیکترین به عنوان نتیجه برمی‌گردد.
یکی از نتایج این روش در زیر مشاهده می‌شود. نتایج کامل در نوتبوک مربوطه موجود است:

```
FastTextEmb().print_similars('ویروس کرونا')
```
Python

```
Output exceeds the size limit. Open the full output data in a text editor
1- title: مراقب افراد بدون علامت باشید / افراد ناقل تا ۴۰درصد قدرت شیوع دارند
1- link: https://www.hidoctor.ir/352514_%d9%85%d8%b1%d8%a7%d9%82%d8%a8-%d8%a7%d9%81%d8%b1%d8%a7%d8%af-%d8%a8%d8%af%d9%88%d9%86-%d8%b9%d9%84%d8%a7%d9%85%d8%aa-%d8%a8%d8%a7%d8%b4%db%8c%d8%af-%d8%a7%d9%81%d8%b1%d8%a7%d8%af-%d9%86%d8%a7%d9%82.html/
------------------------
2- title: نوع جدید ویروس کرونا در ویتنام
2- link: https://www.hidoctor.ir/355062_%d9%86%d9%88%d8%b9-%d8%ac%d8%af%db%8c%d8%af-%d9%88%db%8c%d8%b1%d9%88%d8%b3-%da%a9%d8%b1%d9%88%d9%86%d8%a7-%d8%af%d8%b1-%d9%88%db%8c%d8%aa%d9%86%d8%a7%d9%85.html/
------------------------
3- title: کرونا دلتا دقیقا چیست و چرا خطرناکتر است؟
3- link: https://namnak.com/what-is-delta-covid.p82282
------------------------
4- title: بهترین و تنها راه پیشگیری از ابتلا به کرونا دلتا
4- link: https://namnak.com/covid-19-delta-variant.p82669
------------------------
```

۴- روش transformer:

در این روش نیز ابتدا بر روی داده‌ی موجود پیش‌پردازش‌های لازم از جمله نرمالایز کردن، توکنایز کردن، حذف stop wordها انجام شده است.

این روش از یک مدل pre-trained به نام parsBigBird استفاده می‌کند. این مدل ابتدا دانلود شده و ذخیره می‌شود تا در ادامه از آن استفاده شود.

در ادامه دو مدل train می‌شود: اولی در قسمت vectorize docs type1 وجود دارد. (از آنجایی که parsBigBird در گرفتن وکتورها محدودیت دارد و تنها تعداد مشخصی وکتور می‌تواند بگیرد، در این بخش پاراگراف‌های ۳۰۰ کلمه‌ای به هم متصل شده‌اند. پس از ساختن این داده و ذخیره، در ادامه روش tranformer آن را لود می‌کند تا از آن استفاده کند). در بخش اصلی این قسمت عنوان encode شده و هر یک از پاراگراف‌های ۳۰۰ کلمه‌ای به مدل داده می‌شود و وکتورهای آن‌ها با هم میانگین گرفته می‌شود.

در واقع برای هر سند یک اندیس که شماره‌ی سند است به همراه لیست حاوی یک وکتور برای عنوان و وکتورهای مربوط به پاراگراف‌ها نگهداری می‌شود.

قسمت vectorize docs type2 نیز تقریبا مانند قسمت قبل است با این تفاوت که اولا بر اساس جمله تقسیم می‌شود نه پاراگراف دوما بر خلاف اولی که با tensor میانگین می‌گرفت با نامپای میانگین می‌گیرد.

در قسمت mix to vectors ابتدا یک سری سند غیرقابل استفاده شناسایی می‌شوند. سپس وکتورهایی که در بخش‌های قبل ساخته شده بودند لود می‌شوند و به نسبت ۱ به ۹ عنوان سند و متن سند میانگین گرفته می‌شود. برای مدل اول نسبت ۷ به ۱ استفاده می‌شود. در نهایت نیز سندهای غیرقابل‌استفاده حذف شده و مدل ذخیره می‌شود.

در روش transformer ابتدا مدل‌ها لود شده و کارهای مورد نیاز انجام می‌شود تا سندهای مرتبط با کوئری برگردانده شود.

یکی از نتایج این روش در زیر مشاهده می‌شود. نتایج کامل در نوتبوک مربوطه موجود است:

```
transformer.print_similars('ریزش مو')
```
Python

```
Output exceeds the size limit. Open the full output data in a text editor
1- title: روش جدید برای درمان ریزش مو و کچلی
1- link: https://namnak.com/درمان-کچلی.p31230
------------------------
2- title: دلایل ریزش مو در زنان
2- link: https://namnak.com/ریزش-مو.p65282
------------------------
3- title: شایعترین عوامل موثر در قد کودکان
3- link: https://namnak.com/قد-کودکان.p60301
------------------------
4- title: چرا نگهداری دندان شیری مهم است و فواید ان چیست؟
4- link: https://namnak.com/نگهداری-دندان-شیری.p58074
------------------------
```

**پوشه‌ی health_retrieval:** این پوشه حالت پکیجی است که مدل‌ها را از داخل آن ایمپورت کرده و استفاده می‌کنیم.

**پوشه‌ی models:** در این پوشه تمامی مدل‌ها و اطلاعات مورد نیاز مربوط به آن‌ها ذخیره شده است که در روش‌های مختلف بازیابی، از این پوشه مدل‌های مورد استفاده را لود می‌کنیم.

توجه کنید که در فایل آپلودشده‌ی تمرین، به دلیل حجم بسیار بالای مدل‌ها، آن‌ها را قرار نداده‌ایم و برای ساخته شدن مدل‌ها باید نوت‌بوک‌ها یک بار اجرا شوند تا مدل‌ها ساخته شده و از آن به بعد استفاده شوند.

## ارزیابی MRR:

در این تمرین از ما خواسته شده که برای ارزیابی مدل‌ها از MRR استفاده کنیم. نتایج این ارزیابی برای ۱۰ کوئری به ازای هر مدل در این داک آورده شده است. همانطور که در داک هم دیده می‌شود، نمره‌ی MRR برای هر روش عبارت است از:

- روش boolean: 0.5627
- روش tf-idf: 0.95
- روش fasttext: 0.8387333333
- روش transformer: 0.9276666667

همانطور که انتظار می‌رفت روش boolean از دقت پایینی برخوردار است و لزوما خروجی‌های مرتبطی به ما نمی‌دهد. در نتایج حاصل از مدل‌ها نیز این روش پایین‌ترین MRR را دارد.
همچنین انتظار داشتیم بهترین نتیجه متعلق به transformerها باشد اما از آنجایی که در زبان فارسی مدل pre-trained خیلی خوبی نداریم، اینگونه نشد و بهترین نتیجه ابتدا متعلق به tf-idf و سپس متعلق به transformer است. روش tf-idf با اینکه روشی قدیمی است اما در برخی جاها نتایج مطلوبی می‌دهد و در اینجا نیز نتایج بسیار خوبی از آن حاصل شده است. روش fasttext نیز با اینکه در اینجا به خوبی دو روش دیگر یعنی tf-idf و transformer نشد اما در میان سندهایی که برمی‌گرداند، تعداد زیادی سند مرتبط وجود دارد.
به صورت کلی در این شاخص MRR و با وجود این کوئری‌های داده‌شده tf-idf نتیجه‌ی بسیار خوبی داده است اما ممکن است اگر از شاخص دیگر و یا کوئری‌های دیگری استفاده کنیم چنین نشود.
نکته‌ی قابل توجه دیگر این است که تعداد کوئری‌هایی که در این ارزیابی استفاده شده است بسیار کم بوده است و ممکن است bias داشته باشد. چه بسا با کوئری‌های دیگر، این ترتیب نمره‌های MRR تغییر کند.