گزارش کار: شناسایی نویسنده با استفاده از داده‌های Newsgroups 20

هدف

هدف این پروژه آشنایی با نحوه استفاده از SVM برای دسته‌بندی متون و شناسایی نویسنده است.

توضیحات پروژه

1. **بارگذاری مجموعه داده Newsgroups 20**
2. **پیش‌پردازش داده‌ها) تبدیل متن به ویژگی‌ها با استفاده از(TF-IDF**
3. **آموزش مدل SVM با کرنل خطی**
4. **ارزیابی مدل و گزارش دقت آن**
5. **رسم نمودار ماتریس سردرگمی**

مراحل انجام کار

بارگذاری مجموعه داده Newsgroups 20

- مجموعه داده Newsgroups 20 شامل مجموعه‌ای از مقالات خبری از 20 گروه خبری مختلف است.
- این مجموعه داده با استفاده از تابع `fetch_20newsgroups`از کتابخانه `sklearn.datasets`بارگذاری شد.

پیش‌پردازش داده‌ها

- داده‌های متنی به ویژگی‌های عددی تبدیل شدند با استفاده از `.TfidfVectorizer`
- از توقف کلمات انگلیسی نیز برای بهبود دقت مدل استفاده شد.

تقسیم داده‌ها به مجموعه‌های آموزشی و تست

- داده‌ها به دو مجموعه آموزشی و تست با نسبت 20/80 تقسیم شدند.

آموزش مدل SVM با کرنل خطی

- مدل SVM با کرنل خطی بر روی داده‌های آموزشی آموزش داده شد.

پیش‌بینی و ارزیابی مدل

- مدل بر روی مجموعه داده‌های تست پیش‌بینی انجام داد و دقت مدل محاسبه شد.

رسم ماتریس سردرگمی

- ماتریس سردرگمی برای نمایش عملکرد مدل در تشخیص‌های صحیح و نادرست ترسیم شد.

نتایج

- **دقت مدل** :مدل به دست آمده دقتی معادل 0.92 را نشان داد.
- **ماتریس سردرگمی** :ماتریس سردرگمی نمایش‌دهنده تشخیص‌های صحیح و نادرست مدل بود که به صورت گرافیکی نمایش داده شد.

نتیجه‌گیری

در این پروژه، از مدل SVM با کرنل خطی برای دسته‌بندی متون و شناسایی نویسنده بر اساس داده‌های مجموعه Newsgroups 20 استفاده شد. عملکرد مدل با استفاده از معیار دقت ارزیابی شد و نتایج نشان دادند که مدل SVM در شناسایی نویسنده‌ها عملکرد قابل قبولی دارد. ماتریس سردرگمی نیز نمایشی بصری از عملکرد مدل ارائه داد و تشخیص‌های صحیح و نادرست را نشان داد.