تمرین: تمرین پاک‌سازی داده‌ها

پاکسازی داده‌ها در NLP برای ساخت چت‌بات‌ها ضروری است. داده‌های تمیز به مدل کمک می‌کنند تا بهتر بفهمد و پاسخ دهد. بدون پاکسازی، چت‌بات ممکن است اشتباه کند یا پاسخ‌های بی‌ربط بدهد. در نتیجه، کیفیت چت‌بات به پاکسازی داده‌ها بستگی دارد.

هدف این تمرین، آشنایی با پاک‌سازی داده‌های فارسی است.

- ابتدا فایل QA.xlsx را با Pandas باز کنید.

به تصویر زیر دقت کنید،

	QUESTION	QUESTION
2	سلام، چطور می‌توانم اپلیکیشن بانکواره را دریافت کنم؟	چطور می‌توانم اپلیکیشن بانکواره را دریافت کنم؟
3	خسته نباشید، آیا بدون داشتن کارت ملی جدید می‌توان افتتاح حساب کرد؟	آیا بدون داشتن کارت ملی جدید می‌توان افتتاح حساب کرد؟
4	درود. آیا امکان بازیابی نام کاربری از دست رفته وجود دارد؟	آیا امکان بازیابی نام کاربری از دست رفته وجود دارد؟
5	وقت بخیر! چگونه می‌توان کد پستی را در سامانه تایید کرد؟ ممنون	چگونه می‌توان کد پستی را در سامانه تایید کرد؟
6	کد ملی کاربر: ۱۲۳۴۵۶۷۸۹۱ آیا افتتاح حساب در سامانه رایگان است؟ تشکر	آیا افتتاح حساب در سامانه رایگان است؟
7	کد ملی کاربر: ۲۰۱۸۰۰۱۱۸۴ سلام آیا می‌توان نوع سپرده را در افتتاح حساب انتخاب کرد؟	آیا می‌توان نوع سپرده را در افتتاح حساب انتخاب کرد؟

هدف آن است که متن‌های سمت چپ به متن‌های سمت راست تبدیل شود. یعنی کلماتی مانند (سلام، درود، ممنون و...) باید حذف شوند.

- کد ملی و عدد آن باید با RegEx حذف شود.
- به‌عنوان خروجی، یک فایل اکسل که شامل متن تمیز هست تحویل دهید.

نکات مهم در حل تمرین

- خروجی یک فایل ipynb. داده شود.
- از کامنت گذاری مناسب، سلول بندی، و Markdown استفاده کنید. (بخشی از معیار ارزیابی شما کد نویسی تمیز و خوانایی کد هست.
- یک فایل گزارش متنی جامع و کامل به صورت pdf آماده کنید. مسئله را در آن شرح دهید، راه‌حل خود را نتایج کسب شده، و کدها را با جزئیات کامل توضیح دهید.
- گزارش مرتب و مناسبی ارائه دهید و فونت خوبی در آن استفاده کنید.