**گزارش پیاده سازی مقاله تشخیص کامنت های اسپم در یوتیوب**

استاد درس: دکتر ایوب باقری

دانشجویان : سحر نفیسی 9321170026

علیرضا محمدی 9321170019

دانشگاه کاشان

زمستان 96

# مقدمه

پیاده سازی پیش رو بر اساس مقاله ی Tulio C. Alberto, Comment Spam Filtering on YouTube, IEEE, 2015 انجام شده است. در این مقاله تشخیص اسپم با روش هایی از جمله درخت تصمیم، ماشین بردار پشتیبان، شبکه های بیزین و ... انجام شده است. روشی که ما برای پیاده سازی برگزیدیم روش K نزدیکترین همسایه یا KNN است که با زبان پایتون و به کمک توابع کتابخانه ای nltk در پایتون انجام شده است.

# پیشنیاز های اجرا و توضیحاتی در مورد کد

در اولین قدم، برای اجرای کد نیاز به پایتون نسخه 3 یا بالاتر داریم که از سایت رسمی پایتون قابل دریافت و نصب است. در ادامه لازم است لایبرری هایی که برای پیاده سازی پروژه از آن ها کمک گرفته شده است نصب شود. تنها کتابخانه ی آماده استفاده شده در این پروژه NLTK است که برای پیش پردازش متن کامنت ها (حذف علائم نگارشی، ریشه یابی، حذف کلمات زائد و ...) مورد استفاده قرار گرفته است. برای نصب این کتابخانه کافی است به اینترنت اتصال داشته باشیم و در خط فرمان دستور زیر را اجرا کنیم:

pip install nltk

در ادامه با استفاده از خط فرمان به پوشه ای که فایل main.py در آن قرار دارد می‌رویم و با استفاده از دستور زیر برنامه را اجرا می‌کنیم:

python main.py

اگر دفعه اول است که اسکریپت را اجرا می‌کنید، لازم است ۲ خط زیر را از کامنت خارج کنید:

nltk.download('punkt')

nltk.download('stopwords')

این دو تابع پس از اجرا، به روز ترین لیست از علائم نگارشی و کلمات زائد را از اینترنت دریافت می‌کنند، دفعات بعدی اجرا برای سرعت بیشتر، می‌توانید مجددا این دو خط را کامنت کنید.

# خروجی پیاده سازی و نتایج به دست آمده