غربالگری کلمات feature engineering

ابتدا داده ها ((کلمات)) را براساس ضریب همبستگی که با اسپم بودن دارند یا ندارند ، نمودار هیتمپ، به دو دسته مهم و غیر مهم تقسیم میکنیم . سپس براساس شهودف منطق و دانستن این نکته که ممکن است برخی کلمات مهم بخاطر باقی کلمات و یا کم تکرار بودن ، اشتباه دسته بندی شده باشند باز هم غرباگری انجام میدهیم و در اخر کلمات نهایی را تولید میکنیم.

significant_words :

['all', 'our', 'over', 'remove', 'internet', 'order', 'receive', 'addresses', 'free', 'business', 'email', 'you', 'credit', 'your', '000', 'money',

'hp', 'hpl', 'george', '650', 'labs', '1999', '!', '$', 'longest', 'total', 'spam']

****

insignificant_words :

['make', 'address', '3d', 'mail', 'will', 'people', 'report', 'font', 'lab', 'telnet', '857', 'data', '415', '85', 'technology', 'parts', 'pm',

'direct', 'cs', 'meeting', 'original', 'project', 're', 'edu', 'table', 'conference', ';', '(', '[', '#', 'average']

**********f

alse_significant_words :

['you', 'your', 'george', '650', 'labs', '1999', 'money']

کلمات بی تاثیر به اشتباه با تاثیر شناخته شده : این کلمات در مکالمه های عادی و روزمره افراد نیز به کار می روند و لزوما تاثیر گذار نیستند

false_insignificant_words :

['address', 'telnet', 'technology', 'original', 'edu', 'average', 'font', 'cs', 'project', 'table', '[', '#']

کلمات تاثیر گذار به اشتباه بی تاثیر شناخته شده: این کلمات ببی تاثیر نیستند چون برای پیام های بازرگانی به کار میروند و مردم از براکت ، کمتر استفاده میکنند

****

final_significant_words :

['all', 'our', 'over', 'remove', 'internet', 'order', 'receive', 'addresses', 'free', 'business', 'email', 'credit', '000', 'hp', 'hpl',

'!', '$', 'longest', 'total', 'address', 'telnet', 'technology', 'original', 'edu', 'average', 'font', 'cs', 'project', 'table', '[', '#']

کلمات مهم نهایی غربال شده که 32 تا هستند. دروقع ما 26 کلمه بی تاثیر را از داده 58کلمه ای حذف کردیم که درواقع کمک شایانی به دقت و سرعت میکند

****