При имплементацията на текущия класификатор са използвани методи споменати в секцията „Източници“, като основните подходи са взаимствани от труда [Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora](http://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1217&context=cs_faculty_pubs) на Ron Bekkerman, 2004.

Подготовка на класификацията

1. Премахват се папките които нямат някаква смислена тема.
   * Това са папки които са по подразбиране. Tакива папки в Gmail са:
     1. Trash
     2. Sent Mail
     3. All mail
     4. Spam
     5. Important
     6. Starred

Например, Gmail маркира голяма част от писмата като Important, без да има някакъв видим смисъл – по ското по ключови думи и подател. Обикновено там влизат и писма изпратени като subscription и др. и не бихме искали да тренираме класификатора с тях.

Папки с определена тематика могат да са „News“, „Travel“, “Business”, „Social“, “Projects” и др.

* + Ще премахнем и папки, които имат малък брой писма – под 5. Като по-късно , ако потребителят премести писма в тях може да ги използваме за трениране на класификатора.

1. Трениране на класификатора
   * Тренирането със съществуващите съобщения и тестването ще бъде хронологично, по датата на получаване, т.к. реално това е начинът, по който се натрупват във времето.
   * В много проучвания, по темата с мейлите се използват за feature-и „торба с думи“ от тялото на съобщението, но ние ще разгледаме как се отразява включването на и елементи като Subject, From, Cc, Bcc, комбинирани по различни начини, т.к. ако потребителят е използвал някакъв филтър на писмата в папките, в които се намират, то те са базирани на тези полета.