Versioner af datasæt

**Konstante parametere i datatransformation:**

1. Fjern email-informations tags (fw, re, subject, cc)
2. Erstat specielle tegn eller tokens med strings (NUMBER, DOLLAR, EMAILADRESS, URL)
3. Lav alt lower case
4. Fjern punctuation
5. Tag stammen af alle tokens (stemming)

**Overblik over datasæt versioner:**

* Version 1: Fjern stop-words (standard-sæt)
* Version 2: Fjern få enron-specifikke ting
  1. Fjern enron, ect, com og kaminski (kunne fjerne mange flere men giver ikke mening ifølge Mikkel)
* Version 3: Alle de konstante parametre (se ovenfor)