Trillion features with hashing

Personalized tokens trick

- $\phi_o(token) = \text{hash}(token) \mathbin{\%} 2^b$

- $\phi_u(token) = \text{hash}(u + "\_" + token) \mathbin{\%} 2^b$

- We obtain 16 trillion pairs (user, word) but still 2^b features

text document (email)	bag of words	bag of words (personalized)	$\phi_0(x) + \phi_u(x)$

NEU
Votre
Apotheke
…

NEU
USER123_NEU
Votre
USER123_Votre
Apotheke
USER123_Apotheke
…

1
0
-1
0
0
-1
0
1
0
…