bag of words جدا کردن لغتها از همدیگه

uniqueWords

لغات تکراری حذف میشن و فقط یکتا ها باقی میمونن

numOfWordsA or B

شمردن تعداد تکرار هر لغت در هر سند

TF= Term Frequency

تعداد دفعاتی که یک لغت در سند وجود دارد

تقسیم بر

تعداد کل لغتهای یک سند

IDF= Inverse Data Frequency

نشان می دهد که یک لغت تا چه حدی
در کل سندها پدیدار می شود و
لغتهایی که حایز اهمیت نیستند مانند
..., or the of what
در این دسته قرار میگیرند
هر چه تعداد یک لغت در سند بسیار
زیاد باشد کمتر حایز اهمیت است.

لگاریتم تعداد سندها

تقسیم بر

تعداد سندهایی که شمل لغت مذکور
می باشد.

TF-IDF

لغتهایی که حاوی اطلاعات مهمی هستند را متمایز می کند.

یک وزن به هر لغت داده میشود ، و اصلا ربطی به تعداد تکرار لغت در سند ندارد.

ممکن است لغت کمتر رخ دهد اما کل متن راجع به آن لغت باشد.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents