Term Frequency (TF) as a Function of One Variable to model it using regression (this is simple three papers with “No References”)

Abdelrhman h Elsemary
onnoasd@hotmail.com

Simple Abstract

Text classification has become the focus of attention of many after the increasing interest in the analysis of social media, in this paper we will discuss how to modify the famous TF equation and work with one of its variables as function and model it via regression. Then you can predict future vocabulary that can be used by a particular author.

Keywords: regression, TF, confidence interval

1. Assumption

1) Assume that we have linearity between $\hat{y}_i$ and X_i then:

$$\hat{Y}_i=\beta_0+\beta_1X_i \tag{1}$$

2) Assume that we have these three approaches (may one of them true and other is false in other word in logarithmic logic ‘**if statement with try** ‘):

Let's do this in the case of Text Analysis from Twitter

Y_i is:

1- The place of word we target in tweets (order)
2- The frequency of main word (that’s mean the frequency of the most important word)
3- Frequency of any word we target in all the document (all tweets)

X_i is:

1- Frequency of same word in all the document
2- Frequency of the main word with less importance (second order of importance)
3- Frequency of any word we target in one document (one tweet)

Note that: the first approach in Y_i match with first approach in X_i (The place of word in tweets & Frequency of same word in all the document) and so on.

2. Method

First let's take a closer look at the famous equation of TF:

Let ϕ (τ, δ) be the Term Frequency and Let τ, δ be term, document, where $\tau \in \delta$.

The simplest way to write TF is:

TF = Number of times a term appears in a document / Total number of items in a document.

Let ϕ $(\tau, \delta) = \phi_{\tau,\delta}$

$$TF= \phi_{\tau*,\delta} / \Sigma_{\tau} \phi_{\tau,\delta} \qquad (2)$$

We can describe (2) as: the frequency of the target word in one document / sum of all frequency words in the same document.

Method 1: The strict method is defind as $\phi_{\tau*,\delta}= \mathbf{a}\phi_{\tau,\delta}+\mathbf{b}$ in other word we assume that there is a strict linearity between the frequency of one word and other words frequency in one document.

In practice, the strict method will often not exist.

Method 2: The regression method is defind as $\phi_{\tau*,\delta}= \beta_0 +\beta_1 \phi_{\tau,\delta}$ This means that we have a linear regression relationship between frequency of one word and the frequency of other words in one document.

Theorem: We can write the TF function as a ratio between each document in terms of only $\phi_{\tau*,\delta}$ (our target word) if and only if method 2 are satisfied with significant level.

$$TF_{\delta}= \beta_1\Sigma_{\delta} \phi_{\tau*,\delta} / \Sigma_{\delta} \phi_{\tau*,\delta} - \beta_0 \qquad \text{in term of } \phi_{\tau*,\delta} \qquad (3)$$

proof:

$TF= \phi_{\tau*,\delta} / \Sigma_{\tau} \phi_{\tau,\delta}$ (by definition)

$$TF_{\delta} = \Sigma_{\delta} \phi_{\tau*,\delta} / \Sigma_{\delta,\tau} \phi_{\tau,\delta} \quad \text{over all document} \qquad (4)$$

$$\Sigma_{\delta} \phi_{\tau*,\delta} = \beta_0+ \beta_1\Sigma_{\delta,\tau} \phi_{\tau,\delta} \quad \text{using method 2} \qquad (5)$$

Under a significant level of regression approch in (5) we can write the model in term only of $\Sigma_{\delta} \phi_{\tau*,\delta}$ with constant β_0 and β_1

$TF_{\delta}= \beta_1\Sigma_{\delta} \phi_{\tau*,\delta} / \Sigma_{\delta} \phi_{\tau*,\delta} - \beta_0$ #

Using the previous assumption in 2 we can convert the main functions of TF function to be a couple of arbitrary linear independence to target some of the information we guess is linearly correlated somehow.

Let β_1, β_0 and δ be A, C and ϑ then apply equation (1) in (3) we get:

$$\mathbf{TF_{i,\vartheta} = A\ \Sigma_{\vartheta}\hat{Y}_{i,\vartheta} / \Sigma_{\vartheta}\ \hat{Y}_{i,\vartheta} - C} \qquad \textbf{where } \mathbf{i \rightarrow \{1,3\},\ \vartheta \rightarrow \{1,2, 3,\ldots,n\}} \qquad (6)$$

Find confidence interval (C.I) of $Y_{i,\vartheta}$ and β_1 :

Lemma: Let b_1 is the estimator of the slop of the simple linear regression model in (6), then $b_1 - b_1(0)$ / S.E (b_1) has t distribution with (n-2) degrees of freedom, also can be used to construct 100(1-σ) % then confidence interval for b_1 as follows:

$b_1 \pm t_{1-\sigma/2,\, n-2} * S.E(b_1)$, where $S.E(b_1) = sqrt\ (Var\ (b_1))$ and $Var\ (b_1) = \sigma^2 / S_{xx} = MSE/S_{xx}$, where $S_{xx} = \Sigma\ (X- \bar{X}\ )^2$

C.I for $\hat{Y}_{i,\vartheta}$ in some point i:

$\hat{Y}_{i,\vartheta}\ \wedge \pm t_{1-\sigma/2,\, n-2} * S.E(\hat{Y}_{i,\vartheta})$, where $S.E(\hat{Y}_{i,\vartheta}) = sqrt\ (Var\ (\hat{Y}_{i,\vartheta}))$ (7)

$Var\ (\hat{Y}_{i,\vartheta}) = MSE\ (1/n + [(X_{i,\vartheta} - \bar{x})^2 / S_{xx}])$

C.I for (6) depends on p-value of b_0 if its significant or not, if its significant we will use full equation to write C.I ratio:

$$C.I= A*[\ \hat{Y}_{i,\vartheta} \pm t_{1-\sigma/2,\, n-2} * S.E(\hat{Y}_{i,\vartheta})] / [\ \hat{Y}_{i,\vartheta} \pm t_{1-\sigma/2,\, n-2} * S.E(\hat{Y}_{i,\vartheta})] - C \qquad (8)$$

If b_0 is not significant then:

$$C.I= A*[\ b_1 \pm t_{1-\sigma/2,\, n-2} * S.E(b_1)] / [\ b_1 \pm t_{1-\sigma/2,\, n-2} * S.E(b_1)] - C \qquad (9)$$

3. conclusion

In a simplified way, I would like to find a way to apply equation (1), but in the case of the traditional TF function (2), I will not be able to do this because of the different indexes.

Trying to find a relationship between the frequency of one or more words is a simplified way of writing the function in terms of only one variable.

The regression relationship can be relied on in predicting or in any other uses. The ease of the equation is also a reason for the faster processing of commands in programming languages.

The initial Assumption can be developed for more than 3, just I used these three hypotheses to try to find a good formula that enables me to predict the future words of the author of an article or tweet through his previous tweets.

If the second method is not satisfied, or if R^2 is low, then you cannot pass to the main theory.

This theory in case of insufficient significant level, will suffer from the problem of cumulative error, and to avoid this error must use advanced mathematical methods not mentioned in this paper