# Task6

Nawar Saeed

Artificial Intelligence for the Web, VT21
DT506A

16 april 2021

*Solution.* The goal of BM25 is to be sensitive to term frequency and document length while not adding too many parameters. The formula of BM25 looks like this:

$$BM25 = tf^*.idf$$

where idf is:

$$idf = log_2(\frac{N}{DF})$$

and $tf^*$ is:

$$tf^* = \frac{tf(k+1)}{k\frac{1-b+b.DL}{AVDL} + tf}$$

Moreover let's defined $\alpha$ :

$$\alpha = \frac{1-b+b.DL}{AVDL}$$

Now we want to show that tf * has an upper and lower bound. To make it a little easier, let's ignore the normalization and choose to set $\alpha = 1$. This is possible if and only if b = 0. the term $tf^*$ looks like this now.

$$tf^* = \frac{tf(k+1)}{k\alpha + tf}$$

Lower bound:
Here, the negative values are not so interesting and for that let's set $tf = 0$, in other words let's only check what will happen if a specific word does not occur in the document. It is obtained that $tf^* = \frac{0(k+1)}{kalpha+0} = 0$, meaning that the factor $tf^*$ has a lower bound of 0.
When it comes to upper bound, the limit of tf must be checked, in other word let's analyze what happens when tf goes to infinity. It is obtained:

$$tf^* = \frac{tf(k+1)}{k\alpha + tf} = \frac{k+1}{\frac{k\alpha}{tf} + 1} = k+1$$

and this will happened if $\lim_{tf\to\infty}$ which means that $tf^*$ has an upper bound k + 1.
When $\alpha = 1$, it means that it does not cares about normalization because the formula will looks like $tf^*(k) = \lim_{k\to\infty} \frac{tf(k+1)}{k+tf} \longrightarrow tf^* = tf$. This means that tf is the dominated term.
Now if the alpha becomes something else than 1, tf will no longer be the dominated term,and this means that $\alpha$ can now manipulate and determine normalization. Let's set b = 1 and obtain a full normatization:

$$\alpha = \frac{DL}{AVDL}$$

Putting this in the formula, $tf^*$ will becomes:

$$tf^* = \frac{tf(k+1)}{k\frac{DL}{AVDL} + tf} \qquad (*)$$

let's check now $\lim_{k\to\infty}$, (*) will become:

$$tf^* = \frac{tf(1+\frac{1}{k})}{\frac{DL}{AVDL} + \frac{tf}{k}} \qquad (**)$$

$\lim_{k\to\infty}(\frac{1}{k}) = 0$ and $\lim_{k\to\infty}(\frac{tf}{k}) = 0$ and for this, (**) becomes:

$$tf^* = \frac{tf(1)}{\frac{DL}{AVDL}} = \frac{tfAVDL}{DL} \qquad (***)$$

It means that the total times of occurrence of a term in a document is normilazed based on AVDL and DL. □