Assignment 6

Information Retrieval and Web Search

Ahmed Rekik - 790063 & Tim Henning - 789242

January 10, 2018

1. t-test

1.a. Define a null hypothesis H0 and an alternative hypothesis H1 suited for the described scenario. Do you prefer a one-tailed or two-tailed test?

u0 = 70

H0: u = u0
u = 70

H1: u != u0
u != 70

As the students claim is, that their approach is *exactly* 70% better, this is a two-tailed test.

1.b. Calculate a test statistic t

x̄ = (40+45+50+70+75)/5
= 56

$s = sqrt(V(x))$
$= sqrt((16^2 + 11^2 + 6^2 + 14^2 + 19^2)/5)$
= 13.928

t = (x̄ - u0) / (s / sqrt(n))
= (56 - 70) / (13.928 / sqrt(5))
= -2.247

1.c. Interpret your test result for α = 0.05 with the help of a t-distributions table1. Do the 5 test runs provide enough evidence for the student's claim?

To accept the claim, t has to be greater or equal than t(α; n-1).

t(0.975; 4) = 2.776
t(x) = -t(1-x)
t(0.025; 4) = -2.776

-2.247 >= -2.776

Yes, the 5 test runs provide enough evidence.

2. Ranking with SVMs

2.a. Calculate the optimal separating boundary with its parameters $\vec{w}$ and b.

Witout loosing the generality we define the decision rule as:
if $\vec{w} \cdot \vec{u} + b \geq 0$ **then** $\vec{u}$ is relevant.
Let's define for a relevant document $\vec{x_i} \Rightarrow y_i = \vec{w} \cdot \vec{x_i} + b = +1$
and for an unrelevant document $\vec{x_j} \Rightarrow y_j = \vec{w} \cdot \vec{x_i} + b = -1$

Our goal here is to define the optimal linear separating boundary that decide the relevance of a document $\vec{x_i}$ noted by $\vec{w} \cdot \vec{x_i} + b = y_i$

Therefore we aim to find the two parameters $\vec{w}$ and b that define our separating boundary through our given samples $\vec{x_1}$, $\vec{x_2}$ and $\vec{x_4}$.

We define $\vec{w} = (w_1, w_2)$ and let's assume our SVM equation as : $w_1.x + w_2.y + w_3 = 0$ such that (x, y) are a coordinate of a defined document and w_3 is the constant of the SVM function.

As $\vec{x_1}, \vec{x_2}$ two relevant documents and $\vec{x_4}$ non relevant document such that:
Document $\vec{x_1}$: $w_1.1 + w_2.0 + w_3 = 1$
Document $\vec{x_2}$: $w_1.0 + w_2.1 + w_3 = 1$
Document $\vec{x_4}$: $w_1.0 + w_2.0 + w_3 = -1$
From the third equation, we ensure that $w_3 = -1$
Therefore, as w_3 is already known, $w_1 = 2$ and $w_2 = 2$ hold.
Thereby our SVM optimal separating boundary decision function would be: $w_1.x + w_2.y + w_3 = 2.x + 2.y - 1 =$ **x + y -**$\frac{1}{2}$ $= 0$
From that, we conclude that $\vec{w} = (1, 1)$ and $b = -\frac{1}{2}$, for $\vec{w}.\vec{u} + b = 0$

3. (Programming) Document Embeddings and Optimization

3.a. For at least one phrase query, one boolean query, and one keyword query, report the total processing runtime. In addition to that, report the name of the operation that you identified as the bottleneck (longest runtime).

Table 1: Query Performance

Query	runtime (ms)	bottleneck	runtime (ms)
'european union'	84.96	int_from_base64	68.0
party AND chancellor	162.31	int_from_base64	90.0
negotiate	17.05	int_from_base64	9.0

The bottleneck in our implementation is the method `int_from_base64(x)`. It is used to convert the IDs and position integer values, that were saved as base64, back to real integers. A more efficient binary encoding could solve this.

3.b. Use gensim to learn a vector representation for each comment. For one comment of your choice, find the most similar comment (not the exact same!) and print both.

Chosen comment:

```
  Pretty much all western production is in China.  Russia doesn't produce
anything except raw oil & gas.  Well, it does, but it doesn't have any impact
on the global market.  Russia is still accepting US $ for oil and gas while
China just loves to kiss US bottom.  If they really want to shake US hegemony,
than they should stop acting like US vassals.  The end of petrodollar will
not cause economic collapse of the USA, unless someone makes an larger impact
on their economy.
```

Most similiar comment found by gensim doc2vec:

```
  And the fact that it's everywhere and won't make the pranksters as much
profit.
```