Assignment 6

Information Retrieval and Web Search

Ahmed Rekik - 790063 & Tim Henning - 789242

January 3, 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

$\bar{x}$ = (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 = ($\bar{x}$ - u0) / (s / sqrt(n))
= (56 - 70) / (13.928 / sqrt(5))
= -2.247

1.c. Interpret your test result for $\alpha = 0.05$ 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.

TODO

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.