# Coursework Assignment B - Group 12
# CS4125 - Seminar Research Methodology For Data Science

Nivedita Prasad – 4712099
Aishwarya Shastry – 4743016
Liliana Oliveira – 4767306

March 21, 2018

## Contents



## 1 Part A - Evaluation

1. **What are the best 3 systems for us, and why?**
   Considering that we have multiple measures per system, this is a violation of the independence assumption in which multiple scores from the same subject cannot be regarded as independent from each other. So the scores are going to be rendered *inter-dependently* rather than independent. Random effect for system which allows to resolve this non-independence by assuming a different baseline score for each system. To do this, a concatenation of the three components of a system were put together in a new variable called *System*.

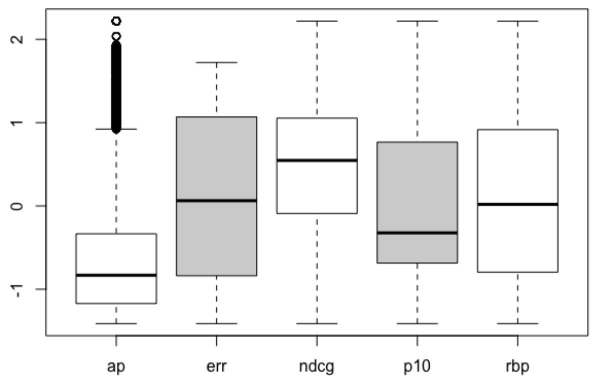


Figure 1: Box Plot of metrics

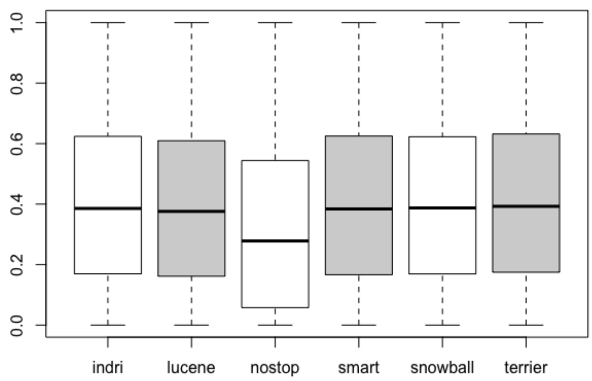


Figure 2: Box Plot of tokens

To assume metric as a fixed-effect we analyze Figure 1 and we wanted to determine if metrics is really having an impact on the general score of a system. Based on this, the following models were created:

**Model 0:** Score $\sim$ ( 1 | System) + (1 | Topic) + (1 | Dataset)

**Model 1:** Score $\sim$ ( 1 | System) + (1 | Topic) + (1 | Dataset) + metric

Since the metrics are not measured in an equal way we need to *standardize* them. The approach considered favors the values that are closer to zero, since the Z-score method was applied.

```
Random effects:
 Groups   Name        Variance  Std.Dev.
 system   (Intercept) 0.0498202 0.22320
 topic    (Intercept) 0.4942549 0.70303
 dataset  (Intercept) 0.0001189 0.01091
 Residual             0.4559667 0.67525
Number of obs: 612000, groups:  system, 612; topic, 200; dataset, 4
```

Figure 3: Random Effects with Topic and System as Random Effects

In Figure 3, it is possible to analyze that there is a clear difference in variance in topic when compared to the variation per system or dataset. The residual shows the random deviations from the predicted values that are *not* due to systems, topics and datasets. After adding metric as a fixed effect the output obtained is displayed in Fig 4 and 5.

```
Random effects:
 Groups   Name        Variance  Std.Dev.
 system   (Intercept) 0.0499516 0.22350
 topic    (Intercept) 0.4942887 0.70306
 dataset  (Intercept) 0.0001279 0.01131
 Residual             0.3242080 0.56939
Number of obs: 612000, groups:  system, 612; topic, 200; dataset, 4
```

Figure 4: Random Effects

```
Fixed effects:
             Estimate Std. Error t value
(Intercept) -0.655406   0.050869   -12.9
metricerr    0.765326   0.002302   332.5
metricndcg   1.111609   0.002302   483.0
metricp10    0.654366   0.002302   284.3
metricrbp    0.745731   0.002302   324.0
```

Figure 5: Fixed effects

In Figure 4 is possible to conclude that even though the variations of system and topic remain the same, by adding a fixed effect, the residual variance slightly decreased.
On Figure 5 it is possible to observe that the coefficients are the slope for the categorical effect of score. The positive values mean that from AP to the others metrics there is a positive difference of maximum 1.11 for NDCG metric. This means that the score is worse when considering AP metric compared to all the others, which was something that was expected as we can see from Figure 1. Then, there's a standard error associated with this slope, and a t-value, which is simply the estimate divided by the standard error.

```
> anova(evaluation.null,evaluation.model)
Data: mydata
Models:
evaluation.null: score ~ (1 | system) + (1 | topic)
evaluation.model: score ~ (1 | system) + (1 | topic) + metric
                 Df     AIC     BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
evaluation.null   4 -317925 -317880 158967  -317933
evaluation.model  8 -526354 -526263 263185  -526370 208436      4  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: Comparison of the two models

After choosing the two models (one considering metric as a fixed effect and other which does not), a significance look (Figure 6) says that metric affects the scores (XI2(4)=208436, $p<0.05$) increasing it by 0.77 ,1.11,0.65 and 0.75 for the following metrics respectively: Err,NDCG,P10 and RBP. This allows us to conclude that the evaluation of a system depends a lot on which metric we consider. In fig 5 we can see that average precision works poorly when compared to other metrics and NDCg also distants itself from the other metrics. So,we propose to remove average precision and NDCg as considered metrics for our system. For best systems, we standardize the score and plot the average score considering only the metrics that have similar estimate on our models and from that see that the new scores close to 0 are the systems with best evaluation.

Table 1: Best Three Systems

| Token | LUG | Model | Score |
|---|---|---|---|
| Lucene | SnowballPorter | dfiz | 0.00595 |
| Lucene | Porter | dfiz | 0.00741 |
| Lucene | Krovertz | dfiz | 0.02074 |

2. **After deployment, our management team is not very happy with the results, and wants our CS department to improve the search engine. Which component should they try to improve, and why?**

To conclude which component should be improved, four models were created.

**Model 0:** Score $\sim$ (1 | Topic) + (1 | Dataset)

**Model 1:** Score $\sim$ lug + (1 | Topic) + (1 | Dataset) + metric

**Model 2:** Score $\sim$ token + (1 | Topic) + (1 | Dataset) + metric

**Model 3:** Score $\sim$ model + (1 | Topic) + (1 | Dataset) + metric

```
> anova(model0,model1,model2,model3)
Data: mydata
Models:
model0: score ~ (1 | topic) + (1 | dataset)
model1: score ~ lug + (1 | topic) + (1 | dataset)
model2: score ~ token + (1 | topic) + (1 | dataset)
model3: score ~ model + (1 | topic) + (1 | dataset)
       Df     AIC     BIC logLik deviance   Chisq Chi Df Pr(>Chisq)
model0  4 -257433 -257388 128721  -257441
model1  9 -258803 -258701 129410  -258821  1379.5      5     <2e-16 ***
model2  9 -271657 -271555 135837  -271675 12853.9      0     <2e-16 ***
model3 20 -270484 -270258 135262  -270524     0.0     11          1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7: Comparison between the 4 models

From the comparison we can conclude that the model which includes the model component has a significant difference from the other models. Which takes us to the conclusion that the model component should be improved.

## 2 Part B - Critical Analysis

We analyzed that most of the flaws are due to presence of few samples, which gave us a small dataset and insufficient topics to evaluate the score. In addition to this, the tokenizer has an impact on LUG and due to that, if a nostop tokenizer is chosen, the LUG will not be able to have a good performance since stemming cannot relate words which have different forms based on grammatical constructs like - "is, am, be" ( which are all stop words ) all represent the same root verb, "be". This makes the processing more complicated. Not only this is a problem but also it cannot relate words that do not have the same prefix, such as better and good should be reduced to a common stem-good. There could be internal validity in the system when the anaylst assessing the topics is tired. We would finally suggest the analysts to be from different backgrounds to get an unbiased selection and accurate difficulty level of the topics. We suggest a metric where relevance score is normalized. They have assumed that documents ranked more than 100 are not relevant which might not be the case. Through multilevel model, we can infer that a system is best for a specific topic because scores have a lot of variance.

# 3 OUTPUT

**OUTPUT of the R code used for Analysis**

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: newscore ~ (1 | system) + (1 | topic) + (1 | dataset)
   Data: mydata

      AIC       BIC    logLik  deviance  df.resid
1260660.4 1260717.0 -630325.2 1260650.4    611995

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5543 -0.7044 -0.0675  0.6929  4.5964

Random effects:
 Groups   Name        Variance  Std.Dev.
 system   (Intercept) 0.0498202 0.22320
 topic    (Intercept) 0.4942549 0.70303
 dataset  (Intercept) 0.0001189 0.01091
 Residual             0.4559667 0.67525
Number of obs: 612000, groups:  system, 612; topic, 200; dataset, 4

Fixed effects:
              Estimate Std. Error t value
(Intercept) -4.255e-11  5.082e-02       0
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: newscore ~ (1 | system) + (1 | topic) + (1 | dataset) + metric
   Data: mydata

      AIC       BIC    logLik  deviance  df.resid
1052231.9 1052333.8 -526106.9 1052213.9    611991

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.3509 -0.6511 -0.0781  0.6319  5.2236

Random effects:
 Groups   Name        Variance  Std.Dev.
 system   (Intercept) 0.0499516 0.22350
 topic    (Intercept) 0.4942887 0.70306
 dataset  (Intercept) 0.0001279 0.01131
 Residual             0.3242080 0.56939
Number of obs: 612000, groups:  system, 612; topic, 200; dataset, 4

Fixed effects:
             Estimate Std. Error t value
(Intercept) -0.655406   0.050869   -12.9
metricerr    0.765326   0.002302   332.5
metricndcg   1.111609   0.002302   483.0
metricp10    0.654366   0.002302   284.3
metricrbp    0.745731   0.002302   324.0

Correlation of Fixed Effects:
           (Intr) mtrcrr mtrcnd mtrc10
metricerr  -0.023
```

```
metricndcg -0.023  0.500
metricp10  -0.023  0.500  0.500
metricrbp  -0.023  0.500  0.500  0.500
Data: mydata
Models:
evaluation.null: newscore ~ (1 | system) + (1 | topic) + (1 | dataset)
evaluation.model: newscore ~ (1 | system) + (1 | topic) + (1 | dataset) + metric
                 Df     AIC     BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
evaluation.null   5 1260660 1260717 -630325  1260650
evaluation.model  9 1052232 1052334 -526107  1052214 208436      4  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Data: mydata
Models:
model0: score ~ (1 | topic) + (1 | dataset)
model1: score ~ lug + (1 | topic) + (1 | dataset)
       Df     AIC     BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
model0  4 -257433 -257388 128721  -257441
model1  9 -258803 -258701 129410  -258821 1379.5      5  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Data: mydata
Models:
model0: score ~ (1 | topic) + (1 | dataset)
model2: score ~ token + (1 | topic) + (1 | dataset)
       Df     AIC     BIC logLik deviance Chisq Chi Df Pr(>Chisq)
model0  4 -257433 -257388 128721  -257441
model2  9 -271657 -271555 135837  -271675 14233      5  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Data: mydata
Models:
model0: score ~ (1 | topic) + (1 | dataset)
model1: score ~ lug + (1 | topic) + (1 | dataset)
model2: score ~ token + (1 | topic) + (1 | dataset)
model3: score ~ model + (1 | topic) + (1 | dataset)
       Df     AIC     BIC logLik deviance   Chisq Chi Df Pr(>Chisq)
model0  4 -257433 -257388 128721  -257441
model1  9 -258803 -258701 129410  -258821  1379.5      5     <2e-16 ***
model2  9 -271657 -271555 135837  -271675 12853.9      0     <2e-16 ***
model3 20 -270484 -270258 135262  -270524     0.0     11          1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Appendices

## A R code

```
#################
#########  SAVE THE ANSWERS IN TXT #######
sink("analysis.txt")
#########  PRINT THE ANSWERS ON THE SCREEN #######
sink(NULL)

#################
install.packages("lme4")
library(lme4)
#####   Read the dataset file   ########
setwd("~/Desktop/SRMDS/AssignmentB/")
mydata <- read.csv(file="data.csv", header=TRUE, sep=",")

# plot the variation of scores per metric
boxplot(mydata$score ~ mydata$metric, col=c("white","lightgray"))
boxplot(mydata$score ~ mydata$token, col=c("white","lightgray"))

mydata$newscore <-  scale(mydata$score)
mydata$newscore <- as.vector(newscore)

boxplot(mydata$newscore ~ mydata$metric, col=c("white","lightgray"))
mydata$system <- paste(mydata$token,mydata$lug,mydata$model)

evaluation.null = lmer(newscore ~ (1|system) + (1|topic) + (1|dataset), data=mydata,REML
evaluation.model = lmer(newscore ~ (1|system) +(1|topic) + (1|dataset) + metric  , data=n
summary(evaluation.null)
summary(evaluation.model)

anova(evaluation.null,evaluation.model)

library(dplyr)

model5<-mydata %>%
  filter(metric=="p10" | metric=="err" | metric=="rbp") %>%
  group_by(lug,token, model) %>%
  summarize(mean_size = mean(newscore))

### Question 2
mydata$system <- NULL
model0 = lmer(score ~  (1|topic) + (1|dataset), data=mydata,REML=FALSE)
model1 = lmer(score ~ lug + (1|topic) + (1|dataset), data=mydata,REML=FALSE)
model2 = lmer(score ~ token + (1|topic) + (1|dataset), data=mydata,REML=FALSE)
model3 = lmer(score ~ model + (1|topic) + (1|dataset), data=mydata,REML=FALSE)
anova(model0,model1)
anova(model0,model2)
anova(model0,model1,model2,model3)
```