Modul 2 - Day 3

Modül-2: Data Cleaning

İçerik: Parsing, Duplicate Elimination, Ensuring Quality: Validity-Accuracy-Completeness, Statistical Analysis, Unix-Linux Terminal

Anahtar sözcükler: Bash, GNU Awk, GNU sed, jq, CSV files, JSON

Araçlar: Shell(s), Excel, R

Data Quality

Kaynak: https://smartbridge.com/data-done-right-6-dimensions-of-data-quality/

1. **Completeness (Tamlık):** Verinin kendi içinde bütünlüğünü temsil eder. (e.g. eksik veri olmaması.)
2. **Consistency (Tutarlılık):** Veri kümeleri arasız tutarsızlık, imbalanced durumlarının olmaması.
3. **Confirmity (Uygunluk):** Her datanın olması gereken formatta tutulması, kolonların veritipleri uygun olmalıdır. (e.g. tarih verisinin tarih formatında olması, ağırlığın integer olması)
4. **Accuracy (Doğruluk):** Verilerin gerçek dünya nesnesini veya açıklanan bir olayı doğru şekilde yansıttığı derecedir.
5. **Integrity (Bütünlük):** Birbirine ilişkili olan iki veri arasında tutarlılık olup olmaması.
6. **Timeliness(Zamanlılık):** Zamanlılık, bilginin beklendiğinde ve ihtiyaç duyulduğunda mevcut olup olmadığını belirtir. Verilerin güncelliği çok önemlidir.

CrispDM: Cross-industry standard process for data mining Not AL foto ekle

Kaynak: https://www.proglobalbusinesssolutions.com/six-steps-in-crisp-dm-the-standard-data-mining-process/

awk ve gnu-sed not al

```
brew install awk;
brew install gnu-sed;
```

```
cat access-log-2.txt | grep robots.txt | wc -l
```

AWK

```
awk '/^42/ {print $0}' access-log-2.txt | wc -l
# 547
```

slash regex'in başı ve sonu

filtre kısmı default işlem print 0'dır

En uzun satır

```
awk '{if (length($0) > max) max=length($0)} END {print max}' access.log.2.txt
```

".awk" uzantılı bir dosya olusturarak

```
awk -f script.awk file-name.txt
```

1000 karakter uzunluğundan fazla olan satırlar

```
awk 'length($0) > 1000' file-name.txt | wc -l
```

$0

$1, $2, $3 sırasıyla boşlukla ayrıldıktan sonraki alanlar

log dosyasında her token bir bölümü ifade ediyor

tail: son 10 satır

head: ilk 10 satır

-n flag ile satır sayısı belirlenebilir

Şubat ayında kaç log düşmüş?

herhangi bir yerde feb yazabilir bakacağımız yer 4. tokende feb yazmalı

```
awk '$4 ~ /Feb/' file-name.txt | wc- l
```

HTTP Kodlarında analiz

```
awk '$9 ~ /200/' file-name.txt | wc -l
```

IP Count

```
{
  freq[$1]++
}

END {
  for (word in freq)
    printf "%d\t%s\n", freq[word],  word
}
```

```
awk '{print $12}' file-name.txt | uniq ;
# Cozulecek awk '{print $12}' file-name.txt | awk '{split($0, a, "/")}' | uniq
```

Mozilla ve Chrome Sayısı

```
awk '{print $12}' file-name.txt | grep 'Chrome' | wc -l
```

netflix datasetinde country birden fazla film cekmıs yonetmenler vs.

csv ve tsv formatlarını ekle

json örnek sitesi: