

07 Tekst, otsing, regulaaravaldised

-

Tekst ja sõne

- Tekst = sümbolite jada mingist tähestikust.
- "Sõne" = formaalne sõna (string), võib olla ka DNA/valgujärjestus.

Kodeering (ASCII, Unicode, UTF-8)

- ASCII = väike ajalooline märgikogum (klassikaliselt 7-bitine).
- Unicode = standard, mis määrab paljude märkide "koodipunktid".
- UTF-8 = üks Unicode'i kodeering: osad märgid võtavad 1 baiti, teised mitu.

Kuidas string arvutis lõpeb

- Variant A: lõputunnus (nt nullbait 0) – eriti C-laadselt.
- Variant B: algusaadress + pikkus (siis lõpp = algus + pikkus).

Infootsing: täpsus ja saagis

- Täpsus (precision) = $TP / (TP + FP)$ → kui "puhtad" on leitud vasted.
- Saagis (recall) = $TP / (TP + FN)$ → kui palju kõigist sobivatest üles leidsid.

Google'i päringute põhiidee

- "fraas" = otsi täpne fraas (sõnad kõrvuti ja samas järjekorras).
- site:domeen = piirab tulemused konkreetsele saidile/domeenile.
- miinus (-sõna) = välista sõna.
- OR = üks või teine (alternatiiv).

Regulaaravaldised (regex) – põhimärgid

- . suvaline märk
- [abc] üks märk hulgast (a või b või c)
 - 0 või rohkem kordust
 - 1 või rohkem kordust
- {m,n} korduste vahemik

- ^ algus, \$ lõpp
- () grupp, | alternatiiv
- \1 = tagasiviide esimesele grupile (praktiline regex-trikk)

grep / egrep kiirelt

- grep muster fail.txt = kuva read, kus muster esineb
- -x = kogu rida peab sobima mustriga
- -w = otsi "tervet sõna"