

Algoritmikerho 2.2.

Juha Harviainen

Puhelinnumerot

- **Ongelma:** Joukko dokumentteja, joista halutaan kerätä kaikki tekstissä esiintyvät puhelinnumerot muotoa +358 DE FGHIJKL
- **Ratkaisu 1:** Käy teksti läpi ja testaa joka sanasta onko puhelinnumero
- Työlästä!
- **Ratkaisu 2:** Regex `\+358 [0-9]{2} [0-9]{7}`

- **Ongelma:** Mitkä englannin sanat alkavat d:llä ja pituus on viidellä jaollinen?
- **Ratkaisu 1:** `word[0] == 'd' and len(word) % 5 == 0`
- **Ratkaisu 2:** Regex `d[a-z]{4}([a-z]{5})*`

General Email Regex (RFC 5322 Official Standard)

```
(?:[a-z0-9!#$%&'*/-?^_`{|}~]+(?:\\. [a-z0-9!#$%&'*/-?^_`{|}~]+)*|"(?:[\\x01-\\x08\\x0b\\x0c\\x0e-\\x1f\\x21\\x23-\\x5b\\x5d-\\x7f]|\\[[\\x01-\\x09\\x0b\\x0c\\x0e-\\x7f]]*)"@(?: (?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?\\. )+[a-z0-9](?:[a-z0-9-]*[a-z0-9])?|\\[(?: (?:25[0-5]|2[0-4][0-9]|01?[0-9])[0-9]?\\. ) {3} (?:25[0-5]|2[0-4][0-9]|01?[0-9])[0-9]?|\\[a-z0-9-]*[a-z0-9]: (?:[\\x01-\\x08\\x0b\\x0c\\x0e-\\x1f\\x21-\\x5a\\x53-\\x7f]|\\[[\\x01-\\x09\\x0b\\x0c\\x0e-\\x7f]]+))\\])
```

Säännölliset lausekkeet (regex)

- Ytimekäs työkalu koodin kirjoittamisen välttelyyn
- Mitkä ovat regexien rakennuspalikat?
- Kuinka rakentaa monimutkaisempi regex?

Merkkijonot

- `kissa` hyväksyy merkkijonon `kissa`
- `kissa|koira` hyväksyy merkkijonot `kissa` ja `kissa`
- Samoin `k(iss|oir)a` – sulut määräävät vaikutusalueen
- `[abc]` on joko `a`, `b` tai `c`
- `[a-d]` on joko `a`, `b`, `c` tai `d`
- `[1-389]` on joko `1`, `2`, `3`, `8` tai `9`

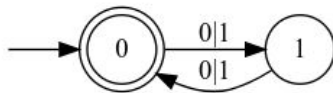
- 1^* sisältää $0-\infty$ merkkiä 1
- $\emptyset^+$ sisältää $1-\infty$ merkkiä $\emptyset$
- $2?$ voi sisältää merkin 2 tai olla sisältämättä
- $(ab)\{3\}$ sisältää kolmesti ab, kun taas $ab\{3\}$ hyväksyy jonon abbb
- $[\emptyset 1]\{3, \}$ sisältää vähintään kolme merkkiä, jotka voivat olla $\emptyset$ tai 1
- $(\emptyset 1?)\{2, 5\}$ sisältää 2-5 nollaa, joista jokaisen perässä saattaa olla 1

Parittomat jonot

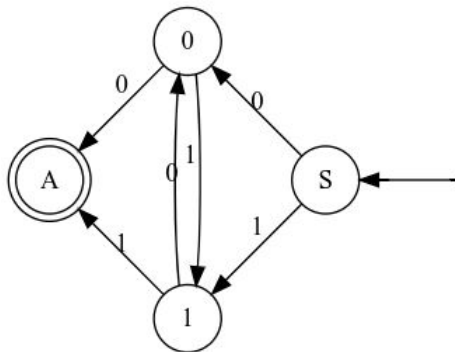
- **Esimerkki:** Anna lauseke, joka hyväksyy parillisen mittaiset merkeistä 0 ja 1 koostuvat merkkijonot
- $(00|01|10|11)^*$
- $((1|0)\{2\})^*$
- $([10][10])^*$
- $([10]\{2\})^*$

Äärelliset automaatit

- Ei välttämättömiä, mutta toisinaan hyödyllisiä
- Parillinen pituus: $((0|1)\{2\})^*$



- Vaihtuva merkki paitsi viimeiset kaksi: $1?(01)^*00|0?(10)^*11$

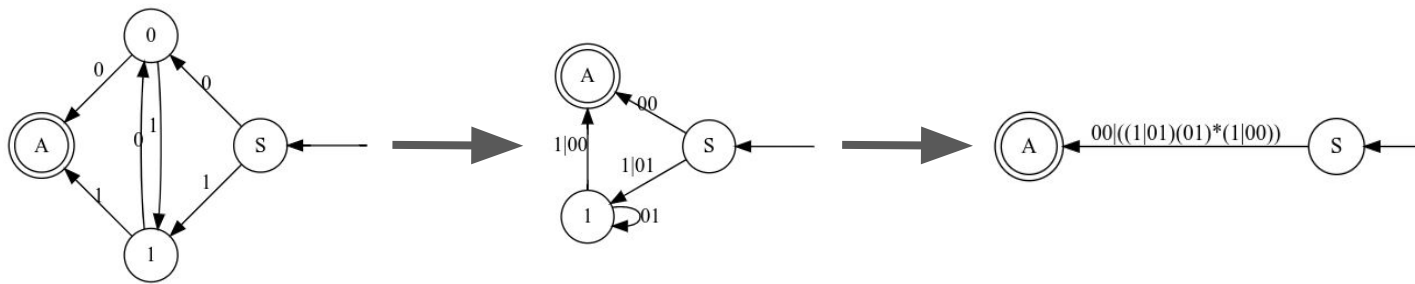


Solmujen eliminointi

- Miten muuntaa automaatti säännölliseksi lausekkeeksi?
- Poistetaan solmut yksitellen!



Viimeiset kaksi samoja



SL-haaste

- Ohjelmointiputkassa SL-haaste säännöllisten lausekkeiden harjoitteluun
- <https://www.ohjelmointiputka.net/slh/>