Verkefnið má finna á GitHub.

1. *Kennitala einstaklinga*

Segðin sem segir til um reglur kennitalna einstaklinga:

\b[0-3]\d(0\d|1[0-2])\d{2}-[2-9]\d{2}[8,9,0]\b

1. \b byrjar segðina.
2. [0-3] þýðir að fyrsta talan í kennitölunni er tala á bilinu 0 upp í 3
3. \d þýðir að næsta tala (önnur tala í kennitölu) má vera hvaða tala sem er á bilinu 0 upp í 9
4. (0\d|1[0-2]) segir til um þriðju og fjórðu tölu í kennitölu. Segði segir að ef þriðja talan er 0 þá getur fjórða talan verið hvaða tala sem er á bilinu 0 upp í 9. Ef þriðja talan er 1 þá getur fjórða talan aðeins verið á bilinu 0 upp í 2 svo mánuður verði löglegur.
5. \d{2} segir svo að næstu tvær tölur geti verið hvaða tölur sem er á bilinu 0 og upp í 9
6. $-$ segir að næst kemur strik, strik á eftir fyrstu 6 tölustöfum.
7. [2-9] segir til um að sjöundi tölustafur er á bilinu 2 og upp í 9 (Þar sem tölustafir 7 og 8 eru alltaf >20).
8. \d{2} segir svo að tölustafir 8 og 9 megi vera hvaða tölustafir sem er.
9. [8,9,0] þýðir að seinasti tölustafurinn er alltaf 8 9 eða 0
10. \b endar segðina.

Kennitala fyrirtækja

Segðin sem segir til um reglur kennitalna fyrirtækja:

\b[4-9]\d{5}-\d{4}\b

1. \b byrjar segðina.
2. Þýðir að fyrst talan þarf að vera á bilinu 4 og upp í 9.
3. Næstu 5 tölur eru á bilinu 0 og upp í 9
4. $-$ þýðir að næst kemur lína. Á eftir fyrstu 6 tölustöfum.
5. \d{4}næstu fjórar tölur eru á bilinu 0 og upp í 9.
6. \b endar svo segðina.

2. *Leit að netfangi*

Hér útfærðum við eina reglulega segð sem leitar eftir löglegum netföngum. Reglulega segðin er eftirfarandi: email_regex = r'(^[a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+.[a-zA-Z]{2,5}.[a-zA-Z]{2,3}$|^[a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+.[a-zA-Z]{2,5}$)'

Hér er svo niðurbrot á því hvað er að gerast í reglulegu segðinni:

“^”:Stendur fyrir upphaf strengs, segðin byrjar að leita frá byrjun strengsins.

“[a-zA-Z0-9_.+-]+”: Er leyfilegir stafir í netfanginu fyrir framan @. Þar mega vera litlir og stórir stafir frá a-z, undirstrik, punktur, plús og bandstrik. Plúsinn í lokin segir að það þarf amk. að vera eitt tákn en þau mega vera fleiri.

“@”: Táknar @ sem er nauðsynlegt fyrir netföng.

[a-zA-Z0-9-]+: Er leyfilegir stafir í netfanginu eftir @. Þar mega vera litlir og stórir stafir frá a-z, tölustafir og bandstrik. Plúsinn í lokin segir að það þarf amk. að vera eitt tákn en þau mega vera fleiri.

.: Punktur er nauðsynlegur til að aðskilja aðal lénið frá top-level domain.

[a-zA-Z]{2,5}: Þetta er fyrir top level domain, það leyfir bókstafi og þarf að vera 2-5 að lengd.

.[a-zA-Z]{2,3}: Þetta er til þess að það sé hægt að hafa undir lén sem er 2-3 stafir.

$: Táknar enda strengsins, segðin leitar að löglegu netfangi og engir aukastafir eru leyfðir á eftir því.

| : Þetta er or/eða aðgerð. Þannig annað hvort þarf fyrri partur segðarinnar að passa eða seinni hlutinn.

^[a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+.[a-zA-Z]{2,5}$: Þetta er síðan seinni parturinn. Sem er eins og fyrri parturinn nema hér eru engin undir lén. En fyrriparturinn var nauðsynlegur fyrir undir lén. Því með því að hafa báðar aðgerðirnar inni fáum við öll netföng sem eru lögleg sem hafa bæði undir lén og þau netföng sem hafa eingöngu top-level lén.

3. *Endurröðun í línum og skrifað CSV -> TSV*

r'([^,]+)\s([^,]+)(?:\s([^,]+))?,\s*([^,]+),\s*([^,]+),\s*(\d+-\d+)'

1. ([^,]+) - Þetta er eiginnafn, og tekur inn allt nema kommu. Samsvarar match.group(1)
2. \s er bilið sem kemur á eftir eiginnafninu
3. ([^,]+) - Þetta er millinafn eða kenninafn. Samsvarar match.group(2)
4. (?:\s([^,]+))?: - "?" merkir að þetta er valfrjálst, en ef það eru þrjú nöfn þá verður þetta kenninafn. Þetta samsvarar match.group(3)
5. , - hérna finnst komma og eftir hana finnst heimilisfangið.
6. \s* - engin eða fleiri bil
7. ([^,]+) - þetta er heimilisfangið og match.group(4)
8. , - komma sem aðskilur heimilisfang frá póstnúmeri
9. \s*([^,]+) - þetta er póstnúmerið og match.group(5).
10. , - komma sem aðskilur póstnúmer og símanúmer.
11. (\d+-\d+) - ein eða fleiri tala, svo bandstrik, svo ein eða fleiri tala. þetta samsvarar match.group(6)

4. *Tímataka*

- byrjun á töfluröð \s* - núll eða fleiri bil eða nýjar línur]*> - [^>]* leyfir hvaða eiginleika sem er innan reitarins nema ">". getur þá verið hvaða klasi sem er. 1 - þetta er reitur sem inniheldur töluna 1 og lokast með <\td>. Fyrsti dálkurinn í röðinni. \s*]*>1 - þessi dálkur inniheldur líka töluna 1. \s* sér um að bil séu meðhöndluð eins og greint var frá að ofan. \s*]*>Simen Nordahl Svendsen - passar við reitinn sem inniheldur textann "Simen Nordahl Svendsen". Finnur röðina þar sem nafnið er "Simen Nordahl Svendsen". (.*?) - passar við hvaða staf sem er nema nýja línu og eins lítið og mögulegt er. Nær innihaldi innan raðarinnar milli nafnsins og lokunarmarkanna .