Matemaattiset ja tilastolliset ohjelmistot, Python-osio Harjoitustyöraportti

Jere Mäkinen
jere.makinen@tuni.fi

toukokuu 2021

Sisällys

1 Johdanto

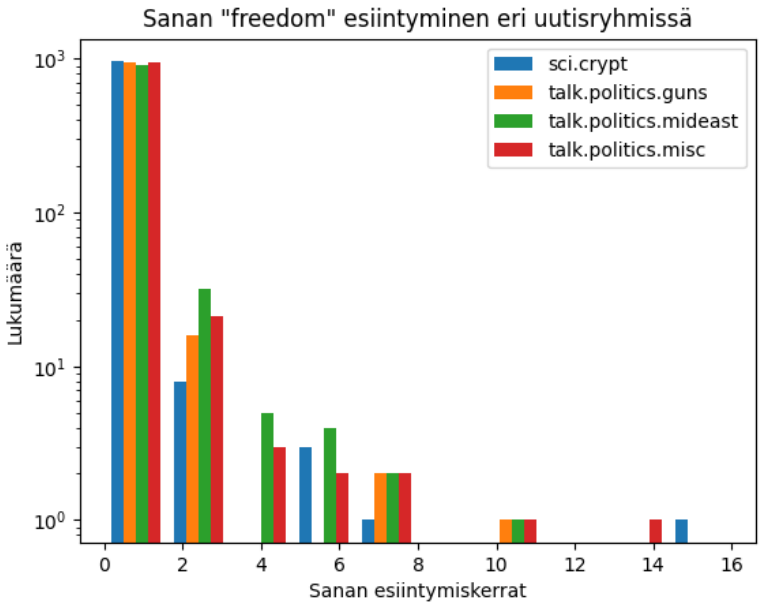
Tässä raportti on harjoitustyön vastausosiona. Raportissa eritellään ratkaisuita tehtävänannon kysymyksiin, sekä esitetään aineistosta piirretyt histogrammit ja kuvaajat. Käsiteltäviä asioita ovat sanojen esiintymismäärät, viestien pituudet, kirjoitusajat, korrelaatiot ja sentimentin anlysointi. Lopuksi vielä yritetään ennustaa uutisryhmiä lineaarista regressiota käyttäen.

Työssä käytetty koodi on .py- ja -txt-tiedostona Moodlen harjoitustyön palautusalueella. Harjoitustyön koodi on tehty PyCharm-kehitysympäristössä.

2 Sanojen esiintymismäärät

Harjoitustyön ensimäisessä osassa tarkastellaan sanojen esiintymismääriä eri uutisryhmissä.

Tarkastellan sanan 'freedom' käyttöä uutisryhmissä sci.crypt, talk.politics.guns talk.politics.mideast ja talk.politics.misc. Sanaa käytettiin eniten ryhmässä talk.politics.mideast (187 kertaa) ja vähiten ryhmässä sci.crypt (101 kertaa). Histogrammi sanan 'freedom' esiintymismääristä on esitetty kuvassa 1.



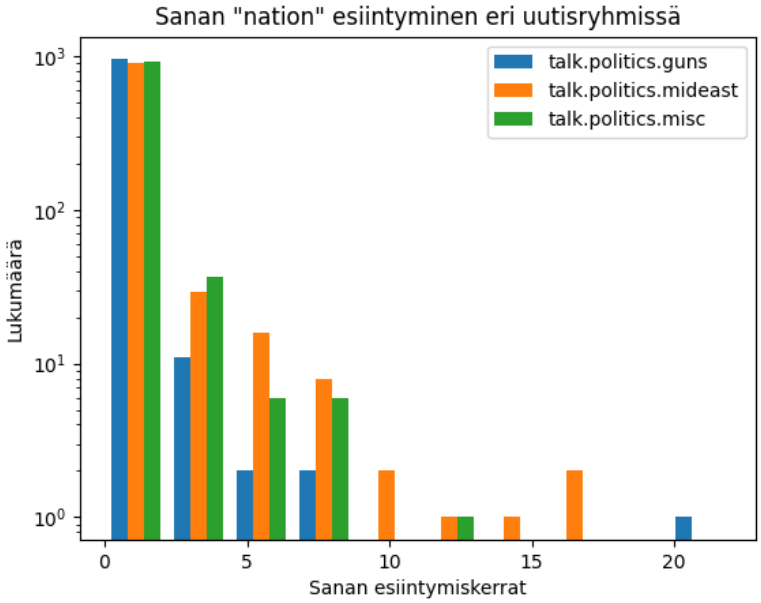
Kuva 1: Sanan 'freedom' käyttö uutisryhmissä sci.crypt, talk.politics.guns talk.politics.mideast ja talk.politics.misc.

Sanan 'freedom' käyttöön liittyviä tunnuslukuja on esitetty taulukossa 1.

Uutisryhmä	Määrä	Keskiarvo	Mediaani	Keskihajonta	0.1% kvantiili	99.9% kvantiili
sci.crypt	101	0.104	0	0.685	0.0	7.24
talk.politics.guns	102	0.105	0	0.5652	0	7.10
talk.politics.mideast	187	0.196	0	0.766	0.0	7.14
talk.politics.misc	148	0.151	0	0.804	0.0	10.09

Taulukko 1: Sanan 'freedom' käytön jakauma eri uutisryhmissä.

Tarkastellan seuraavaksi sanan 'nation' käyttöä uutisryhmissä talk.politics.guns, talk.politics.mideast ja talk.politics.misc. Sanaa käytettiin eniten ryhmässä talk.politics.mideast (527 kertaa) ja vähiten ryhmässä talk.politics.guns (231 kertaa). Histogrammi sanan 'freedom' esiintymismääristä on esitetty kuvassa 2.



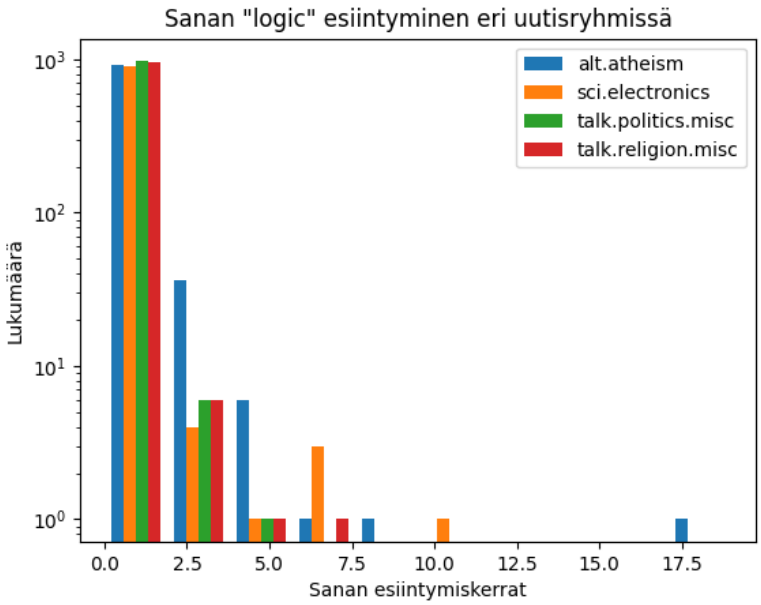
Kuva 2: Sanan 'nation' käyttö uutisryhmissä talk.politics.guns, talk.politics.mideast ja talk.politics.misc.

Sanan 'nation' käyttöön liittyviä tunnuslukuja on esitetty taulukossa 2.

Uutisryhmä	Määrä	Keskiarvo	Mediaani	Keskihajonta	0.1% kvantiili	99.9% kvantiili
talk.politics.guns	231	0.238	0	0.965	0.0	7.48
talk.politics.mideast	527	0.551	0	1.56	0.0	17.0
talk.politics.misc	373	0.381	0	1.077	0.0	8.069

Taulukko 2: Sanan 'nation' käytön jakauma eri uutisryhmissä.

Tarkastellan seuraavaksi sanan ’logic’ käyttöä uutisryhmissä alt.atheism, sci.electronics, talk.politics.misc ja talk.religion.misc. Sanaa käytettiin eniten ryhmässä alt.atheism (205 kertaa) ja vähiten ryhmässä talk.politics.misc (45 kertaa). Histogrammi sanan ’logic’ esiintymismääristä on esitetty kuvassa 3.



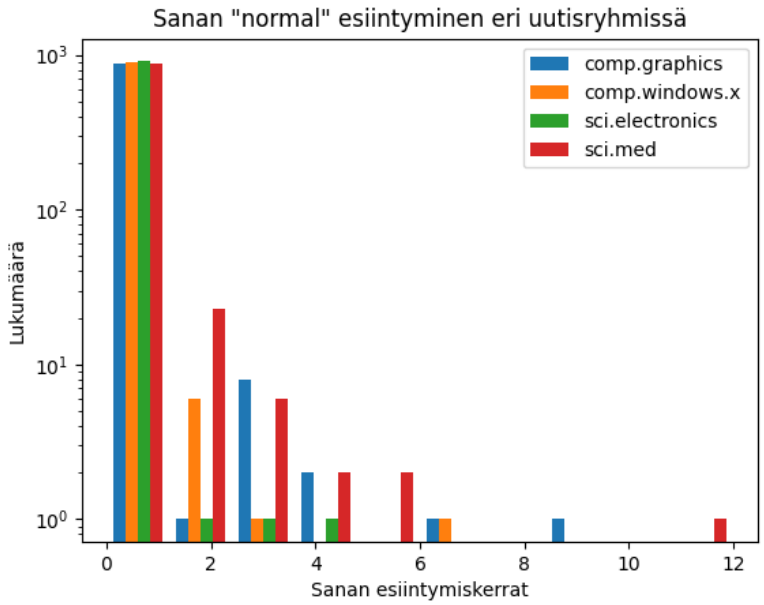
Kuva 3: Sanan ’logic’ käyttö uutisryhmissä alt.atheism, sci.electronics, talk.politics.misc ja talk.religion.misc.

Sanan ’logic’ käyttöön liittyviä tunnuslukuja on esitetty taulukossa 3.

Uutisryhmä	Määrä	Keskiarvo	Mediaani	Keskihajonta	0.1% kvantiili	99.9% kvantiili
alt.atheism	205	0.210	0	0.901	0.0	8.297
sci.electronics	66	0.0726	0	0.560	0.0	7.276
talk.politics.misc	45	0.0460	0	0.284	0.0	3.023
talk.religion.misc	58	0.0604	0	0.360	0.0	5.040

Taulukko 3: Sanan ’logic’ käytön jakauma eri uutisryhmissä.

Tarkastellan seuraavaksi sanan ’normal’ käyttöä uutisryhmissä comp.graphics, comp.windows.x, sci.electronics ja sci.med. Sanaa käytettiin eniten ryhmässä sci.med (133 kertaa) ja vähiten ryhmässä sci.electronics (35 kertaa). Histogrammi sanan ’normal’ esiintymismääristä on esitetty kuvassa 4.



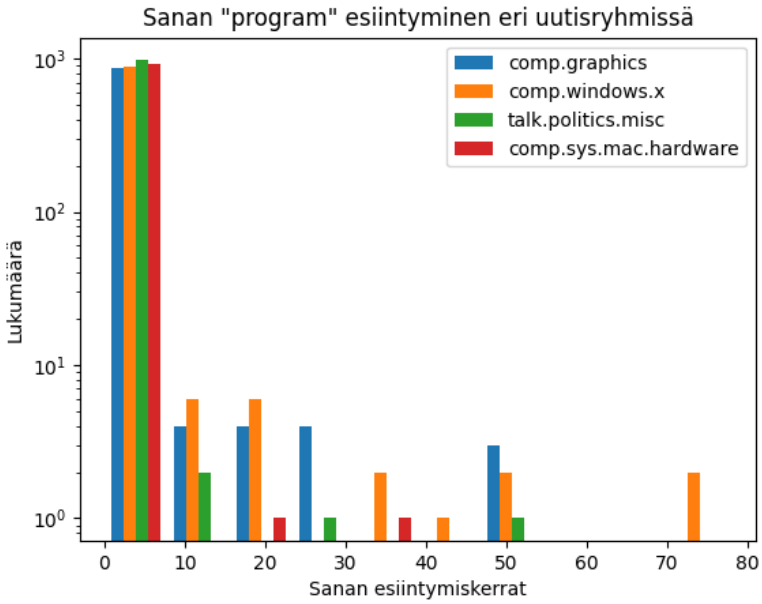
Kuva 4: Sanan ’normal’ käyttö uutisryhmissä comp.graphics, comp.windows.x, sci.electronics ja sci.med.

Sanan ’normal’ käyttöön liittyviä tunnuslukuja on esitetty taulukossa 5.

Uutisryhmä	Määrä	Keskiarvo	Mediaani	Keskihajonta	0.1% kvantiili	99.9% kvantiili
comp.graphics	62	0.0696	0	0.512	0.0	6.330
comp.windows.x	51	0.0562	0	0.326	0.0	3.282
sci.electronics	35	0.0385	0	0.243	0.0	3.092
sci.med	133	0.145	0	0.655	0.0	5.602

Taulukko 4: Sanan ’normal’ käytön jakauma eri uutisryhmissä.

Tarkastellan vielä lopuksi sanan 'program' käyttöä uutisryhmissä comp.graphics, comp.windows.x, talk.politics.misc, ja comp.sys.mac.hardware. Sanaa käytettiin eniten ryhmässä comp.windows.x (914 kertaa) ja vähiten ryhmässä comp.sys.mac.hardware (123 kertaa). Histogrammi sanan 'normal' esiintymismääristä on esitetty kuvassa 5.



Kuva 5: Sanan 'program' käyttö uutisryhmissä comp.graphics, comp.windows.x, talk.politics.misc, ja comp.sys.mac.hardware.

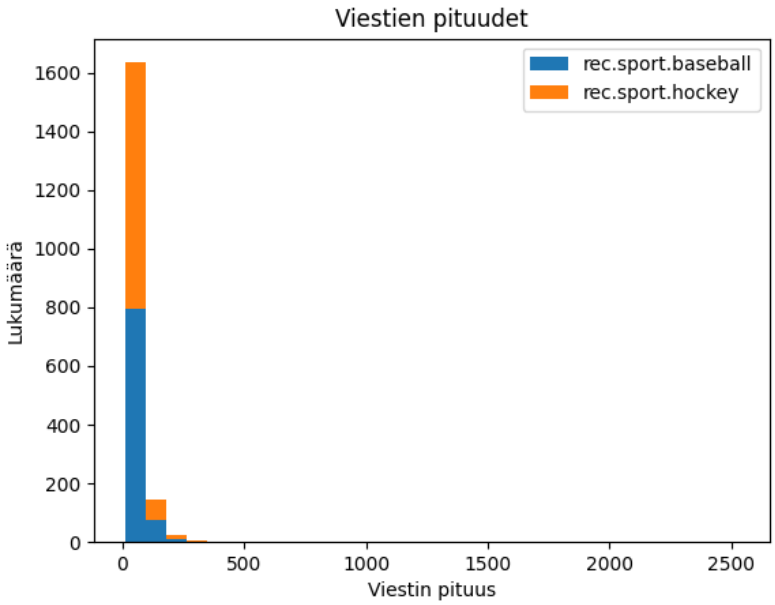
Sanan 'program' käyttöön liittyviä tunnuslukuja on esitetty taulukossa 5.

Uutisryhmä	Määrä	Keskiarvo	Mediaani	Keskihajonta	0.1% kvantiili	99.9% kvantiili
comp.graphics	812	0.911	0	3.84	0.0	47.0
comp.windows.x	914	1.008	0	5.223	0.0	78.0
talk.politics.misc	272	0.278	0	2.080	0.0	30.48
comp.sys.mac.hardware	123	0.134	0	1.302	0.0	19.3

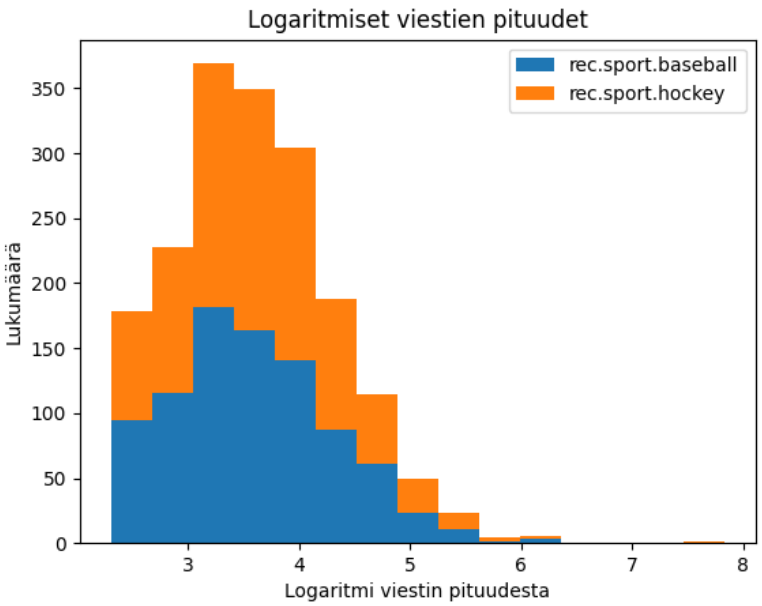
Taulukko 5: Sanan 'program' käytön jakauma eri uutisryhmissä.

3 Viestien pituudet

Kuvassa 6 on esitetty viestien pituudet uuritsryhmissä rec.sport.baseball ja rec.sport.hockey. Kuvassa 7 puolestaan on viestien pituudet logaritmi muutoksen jälkeen. Näistä voidaan huomata, että pituuksien logaritmien jakauma muistuttaa huomattavasti enemmän normaalijakaumaa.



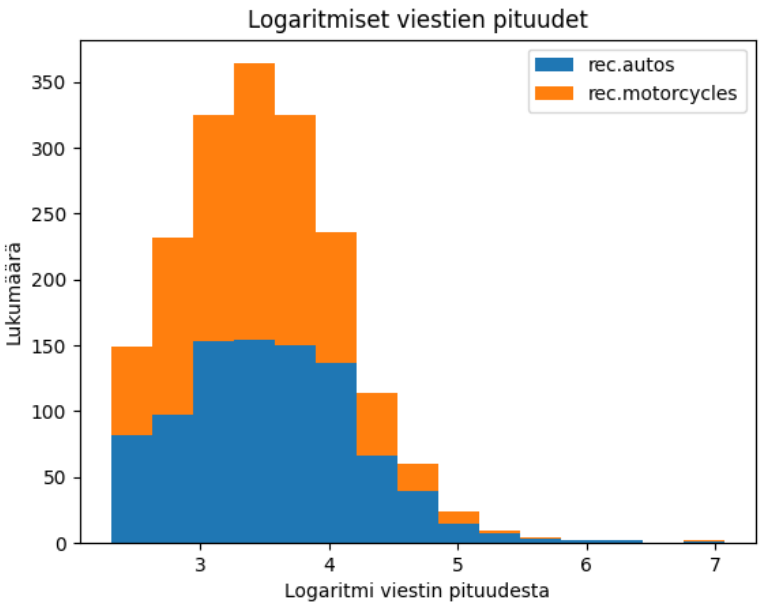
Kuva 6: Viestien pituudet uutisryhmissä rec.sport.baseball ja rec.sport.hockey.



Kuva 7: Viestien pituudet uutisryhmissä rec.sport.baseball ja rec.sport.hockey logaritmimuunnoksen jälkeen.

Kun suoritetaan odotusarvojen t-testi ryhmille rec.sport.baseball ja rec.sport.hockey, saadaan testin palauttamaksi p-arvoksi $0,247$. Tämän perusteella ei voida hylätä hypoteesia siitä, että logaritmisten viestien pituuksien jakaumat näiden ryhmien välillä olisivat erilaiset.

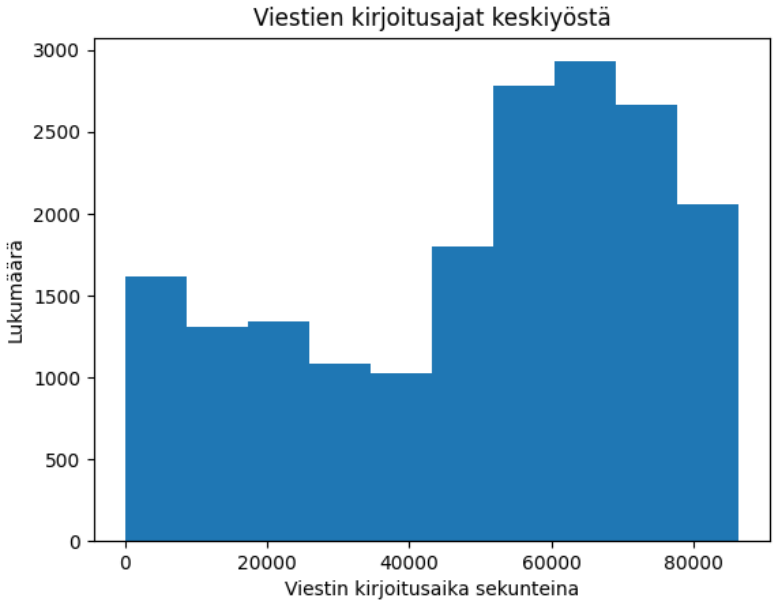
Samaisen testin tulos uutisryhmille rec.autos ja rec.motorcycles on $6.07 \cdot 10^{-5}$. Näiden kohdalla siis voidaan sanoa, että logaritmisten viestien pituuksien jakaumat eivät mitä todennäköisimmin ole samat. Jakaumien erovaisuutta voidaan huomata myös kuvasta 8. Tosin koska viestien pituudet eivät logaritmimuunnoksen jälkeenkään ole tarkalleen normaalijakautuneita, voidaan molempien testien tuloksia pitää korkeintaankin suuntaa antavina.



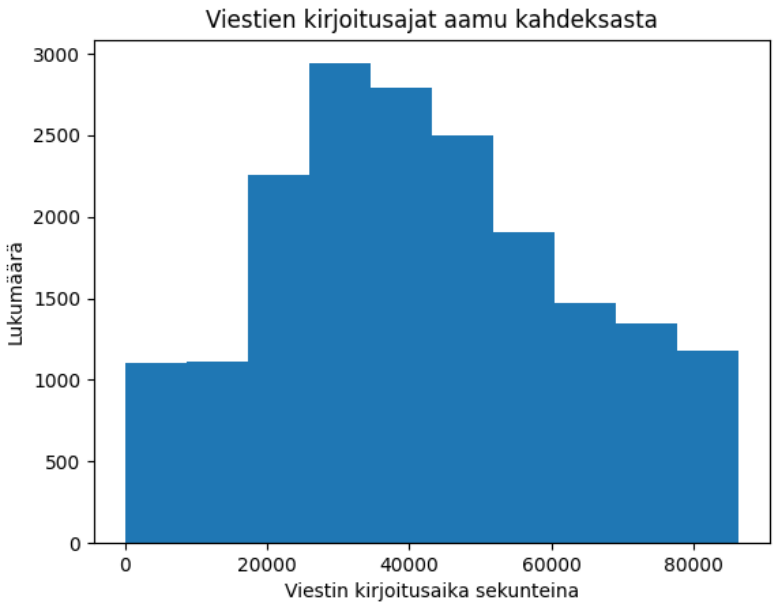
Kuva 8: Viestien pituudet uutisryhmissä rec.autos ja rec.motorcycles logaritmimuunnoksen jälkeen.

4 Kirjoitusajat

Kuvissa 9 on esitetty viestien kirjoitusajat keskiyöstä alkaen. Kuvassa 10 puolestaan on kirjoitusajat muokattuna alkamaan aamu kahdeksasta. Kuvista huomataan, että muokkauksen jälkeen viestien kirjoitusajat muistuttavat huomattavasti enemmän normaalijakaumaa.



Kuva 9: Viestien kirjoitusajat.



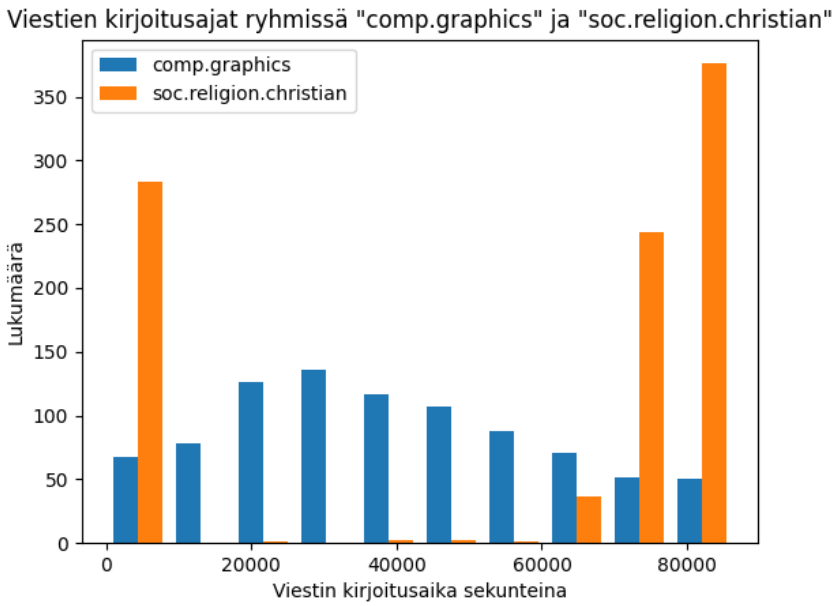
Kuva 10: Viestien kirjoitusajat muokattuna

Muokattujen kirjoitusaikojen keskiarvo, mediaani ja keskihajonta on esitetty taulukossa 6

Keskiarvo	Mediaani	Keskihajonta
42041.1	40327.5	21186.9

Taulukko 6: Muokatun kirjoitusajan tunnuslukuja.

Tarkastellaan vielä erikseen kirjoitusaikoja uutisryhmissä comp.graphics ja soc.religion.christian. Odotusarvojen t-testi antaa näiden ryhmien jakaumien kohdalla tuloksen $1,503 \cdot 10^{-33}$. Jakaumien välillä on tämän testin perusteella siis merkitsevä ero. Sama voidaan myös päätellä kuvasta 11.



Kuva 11: Viestien muokatut kirjoitusajat ryhmissä comp.graphics ja soc.religion.christian.

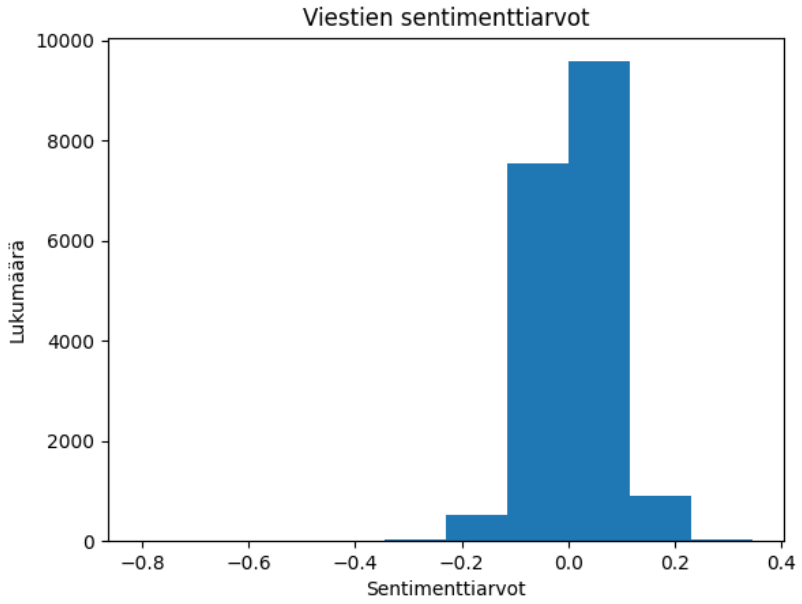
5 Korrelaatiot

Korrelaatiokerroin sanojen 'jpeg' ja 'gif' välillä koko aineistossa on $0,966$. Sanojen välinen korrelaatio on siis hyvinkin suurta. Sanojen 'sale' ja 'write' kohdalla puolestaan korrelaatio ei ole niin merkittävää, sillä korrelaatiokerroin on -0.082.

Sanojen 'jpeg' ja 'gif' välinen korrelaatio on vielä muuta aineistoa surempaa uutisryhmässä comp.graphics. Korrelaatiokerroin tässä ryhmässä on $0,991$.

6 Sentimentin analysointi

D'Agostinon ja Pearsonin testin perusteella sentimenttiarvo ei ole normaalijakautunut yli koko aineiston, sillä testin antama p-arvo on 0,0. Koko aineiston senttimenttijakauma on esitetty kuvassa 12.

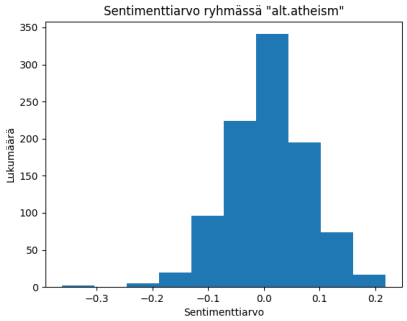


Kuva 12: Sentimenttijakauma koko aineistossa.

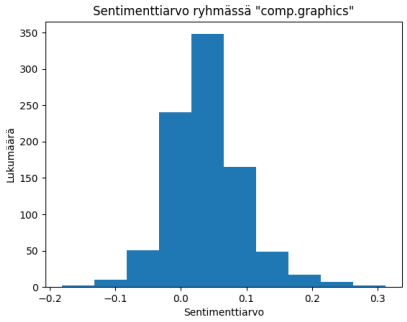
Sentimentin jakaumat eri uutisryhmissä on esitetty taulukossa 7 ja kuvissa 13-32.

Uutisryhmä	Keskiarvo	Mediaani	Keskihajonta	25% kvantiili	75% kvantiili
alt.atheism	0.00865	0.0102	0.0725	-0.0337	0.0530
comp.graphics	0.0387	0.0357	0.057369	0.0	0.0690
comp.os.ms-windows.misc	0.0254	0.0309	0.0822	0.0	0.0606
comp.sys.ibm.pc.hardware	0.0216	0.0207	0.05940	-0.00845	0.0541
comp.sys.mac.hardware	0.0308	0.0299	0.0645	0.0	0.0676
comp.windows.x	0.0276	0.0246	0.0586	0.0	0.0593
misc.forsale	0.0400	0.0382	0.0537	0.0	0.0724
rec.autos	0.00957	0.0152	0.0593	-0.0257	0.0468
rec.motorcycles	0.00130	0.0	0.0592	-0.0339	0.0370
rec.sport.baseball	0.0262	0.0270	0.0605	-0.00232	0.0608
rec.sport.hockey	0.0259	0.0234	0.0582	-0.00652	0.0592
sci.crypt	0.0170	0.0157	0.0598	-0.0196	0.0550
sci.electronics	0.0357	0.0333	0.0551	0.0	0.0722
sci.med	-0.00450	0.0	0.0715	-0.0465	0.0401
sci.space	0.0104	0.0118	0.0557	-0.0254	0.0431
soc.religion.christian	0.0237	0.0278	0.0779	-0.0217	0.0659
talk.politics.guns	-0.02043	-0.0196	0.0658	-0.0625	0.0206
talk.politics.mideast	-0.0217	-0.0216	0.0645	-0.0607	0.0157
talk.politics.misc	-0.0111	-0.00966	0.06406	-0.0509	0.0286
talk.religion.misc	0.00550	0.00408	0.0698	-0.0370	0.0496

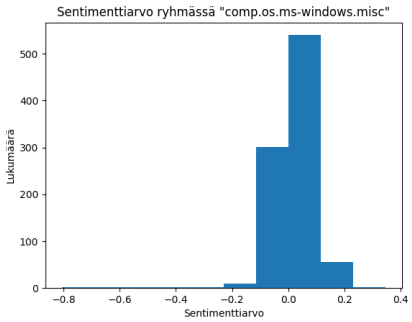
Taulukko 7: Sentimentti arvojen tunnuslukuja.



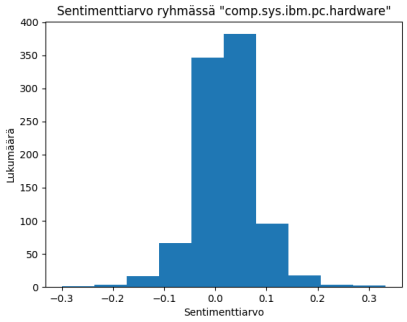
Kuva 13: Sentimenttijakauma ryhmässä alt.atheism.



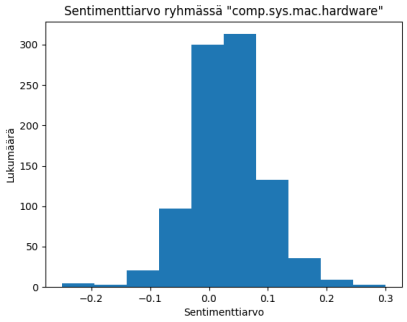
Kuva 14: Sentimenttijakauma ryhmässä comp.graphics.



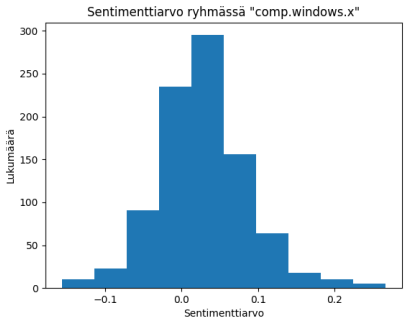
Kuva 15: Sentimenttijakauma ryhmässä comp.os.ms-windows.misc.



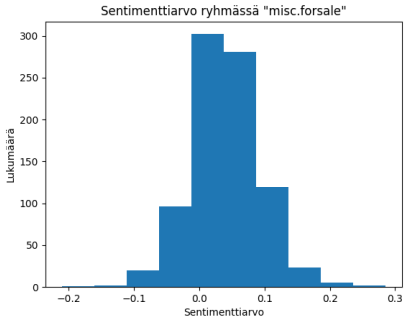
Kuva 16: Sentimenttijakauma ryhmässä comp.sys.ibm.pc.hardware.



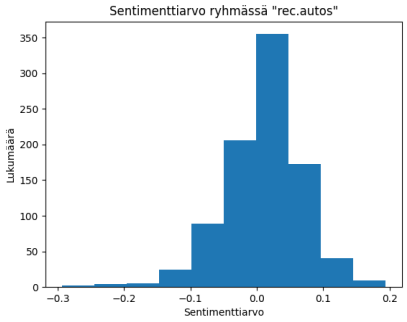
Kuva 17: Sentimenttijakauma ryhmässä comp.sys.mac.hardware.misc.



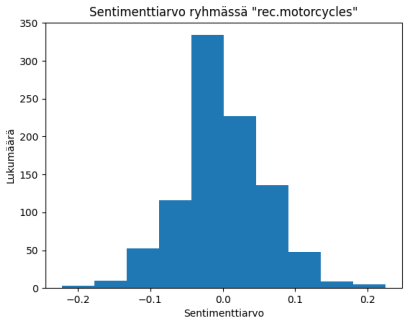
Kuva 18: Sentimenttijakauma ryhmässä comp.windows.x.



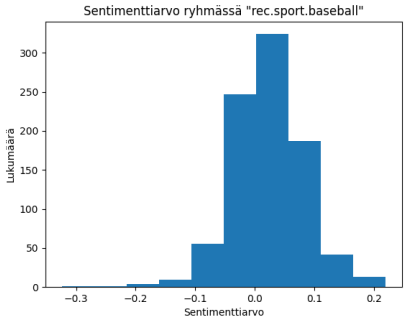
Kuva 19: Sentimenttijakauma ryhmässä misc.forsale.



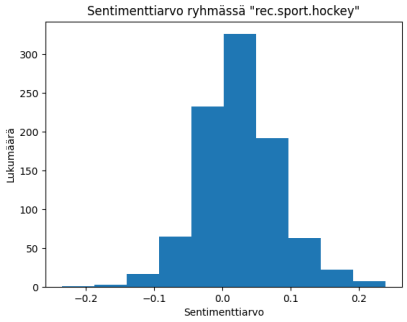
Kuva 20: Sentimenttijakauma ryhmässä rec.autos.



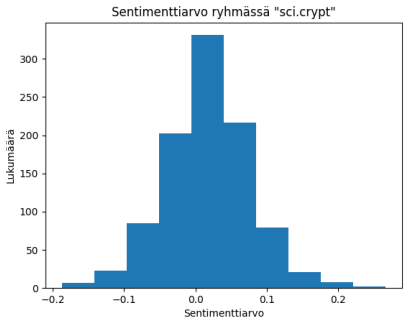
Kuva 21: Sentimenttijakauma ryhmässä rec.motorcycles.



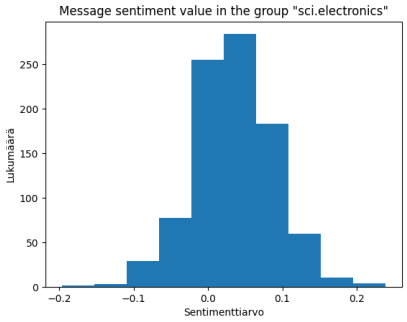
Kuva 22: Sentimenttijakauma ryhmässä rec.sport.baseball.



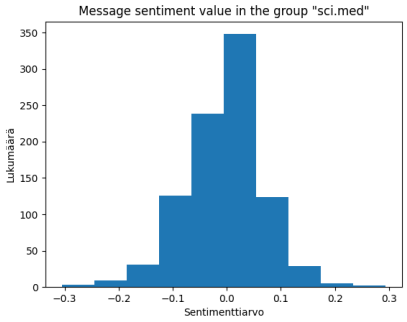
Kuva 23: Sentimenttijakauma ryhmässä rec.sport.hockey.



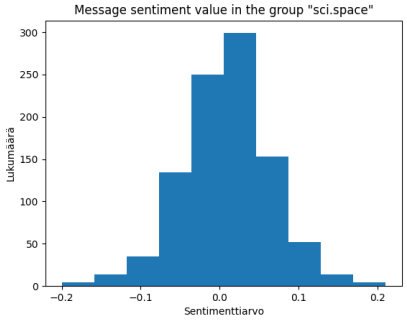
Kuva 24: Sentimenttijakauma ryhmässä sci.crypt.



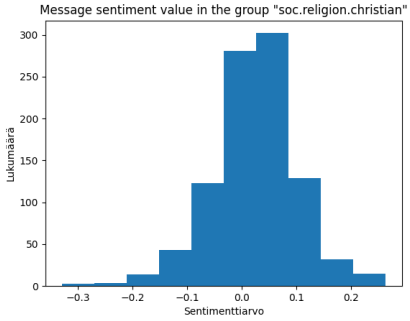
Kuva 25: Sentimenttijakauma ryhmässä sci.electronics



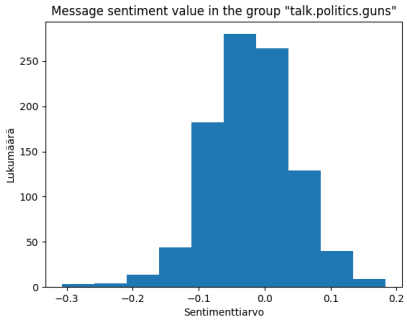
Kuva 26: Sentimenttijakauma ryhmässä sci.med



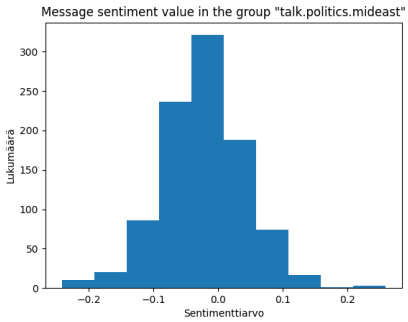
Kuva 27: Sentimenttijakauma ryhmässä sci.space.



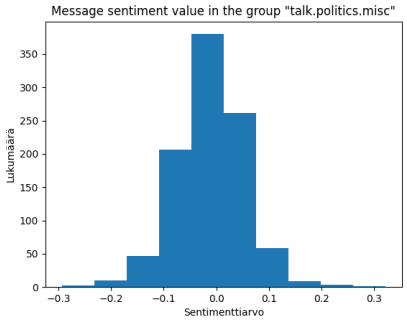
Kuva 28: Sentimenttijakauma ryhmässä soc.religion.christian.



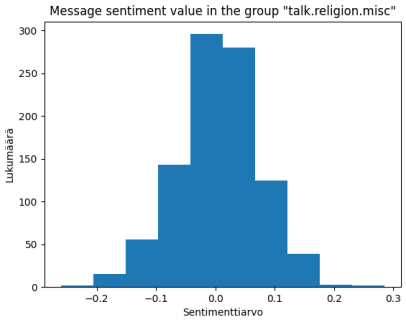
Kuva 29: Sentimenttijakauma ryhmässä talk.politics.guns.



Kuva 30: Sentimenttijakauma ryhmässä talk.politics.mideast.



Kuva 31: Sentimenttijakauma ryhmässä talk.politics.misc.



Kuva 32: Sentimenttijakauma ryhmässä talk.religion.misc.

Kolme positiivisinta uutisryhmää (suluissa keskiarvo) ovat:

1. misc.forsale (0.0400)
2. comp.graphics (0.0387)
3. sci.electronics (0.0357)

Kolme negatiivisinta puolestaan ovat:

1. talk.politics.mideast (-0.0217)
2. talk.politics.guns (-0.0204)
3. talk.politics.misc (-0.0111)

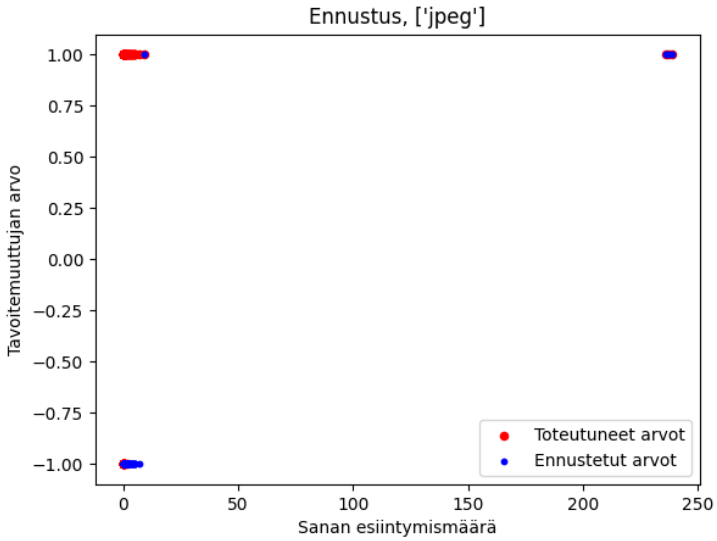
Odotusarvojen t-testin mukaan ryhmien comp.sys.ibm.pc.hardware ja comp.sys.mac.hardware ja rec.autos ja rec.motorcycles välillä sentimenttijakaumissa on eroja. Testin p-arvot ovat vastaavasti 0.00143 ja 0.00273. Myös tutkittaessa uutisryhmä kohtaisia tunnuslukuja taulukosta 7, voidaan huomata merkittäviä eroja näiden ryhmien kohdalla.

Sen sijaan ryhmien rec.sport.baseball ja rec.sport.hockey sentimenttijakaumat voivat testin perusteella olla samanlaiset, sillä testin antama p-arvo on 0.898. Myös taulukon 7 luvut, ovat näiden ryhmien kohdalla melko lähellä toisiaan. Nämäkin tulokset ovat kuitenkin korkeintaankin suuntaa antavia, sillä kuten aiemmin todettiin, eivät sentimenttiarvot ainakaan koko aineistossa todennäköisesti ole tarkalleen normaalijakautuneita.

7 Uutisryhmän ennustaminen

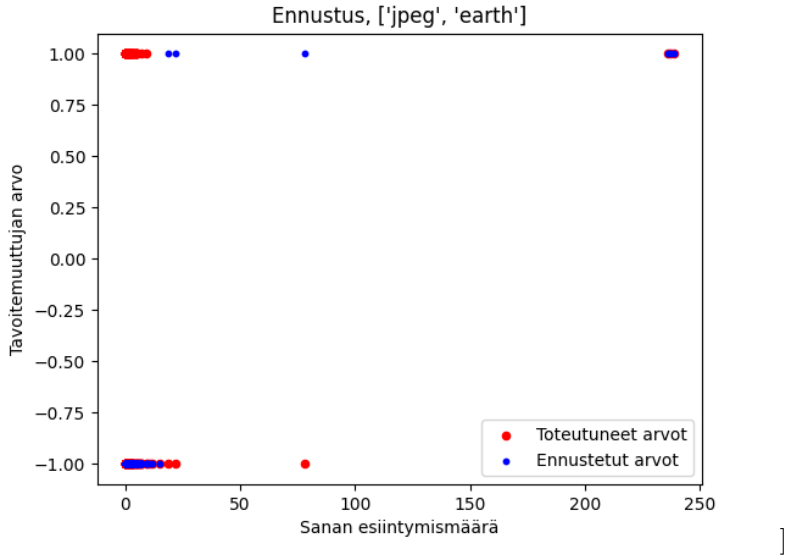
Tarkastellaan viestejä uutisryhmissä comp.graphics ja sci.space. Annetaan jokaiselle viestille tavoitemuuttujan arvo 1, jos se kuuluu ryhmään comp.graphics ja -1 jos se kuuluu ryhmään sci.space. Seuraavaksi yritetään ennustaa lineaarisen regression avulla kumpaan ryhmään viesti kuuluu, hyödyntäen tiettyjen sanojen esiintymismääriä. Ennusteen antamat tavoite muuttujan arvot on kuvia varten pyöristetty vastaamaan joko arvo -1 tai 1. Ennustetta verrataan toteutuneisiin arvoihin piirtämällä ennusteen pisteet samaan kuvaan toteutuneiden kanssa. Mitä paremmin pisteet osuvat päällekäin, sen parempi ennustus on.

Ensimmäiseksi käytetään vain yhtä sanaa: 'jpeg'. Tämän ennustuksen keskimääräinen neliövirhe on 0,9963. Ennustuksen tulokset on esitetty kuvassa 33.



Kuva 33: Ennuste sanan 'jpeg' esiintymismäärän perusteella

Seuraavaksi tehdään ennustus käyttäen sanoja 'jpeg' ja 'earth'. Tämän ennusteen keskimääräinen neliövirhe on 0,998. Ennustuksen tulostukset on esitetty kuvassa 34.



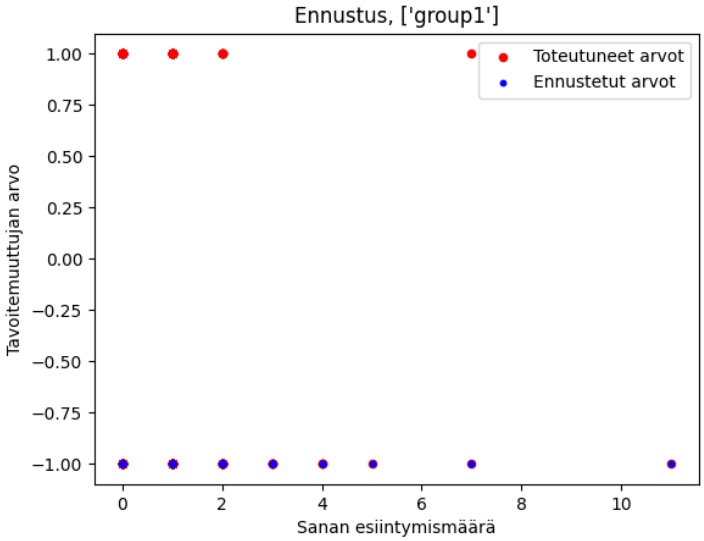
Kuva 34: Ennuste sanojen ’jpeg’ ja ’earth’ esiintymismäärän perusteella

Yritetään nyt ennustusta ennalta määriteltyjen kahdeksan sanaryhmän pohjalta. Keskimääräiset neliövirheet on esitetty taulukossa 8.

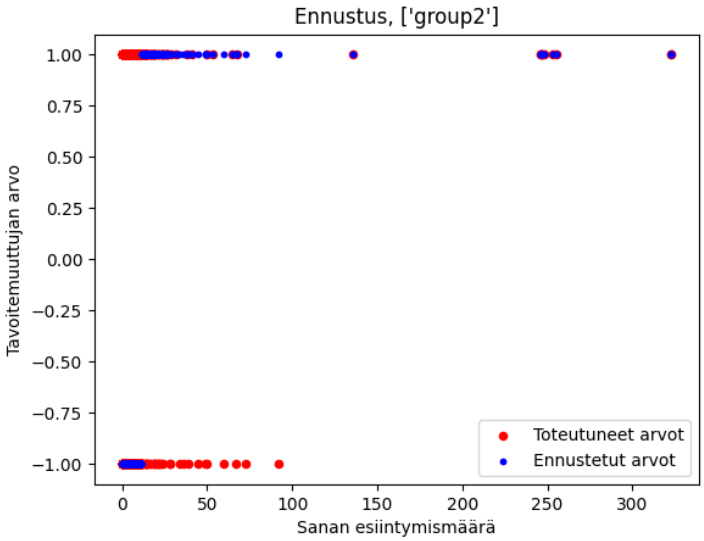
Sanaryhmä	Keskimääräinen neliövirhe
Ryhmä 1	0.990
Ryhmä 2	0.987
Ryhmä 3	0.974
Ryhmä 4	0.997
Ryhmä 5	0.998
Ryhmä 6	0.996
Ryhmä 7	0.985
Ryhmä 8	0.999

Taulukko 8: Keskimääräiset neliövirheet ennustuksille eri sanaryhmissä.

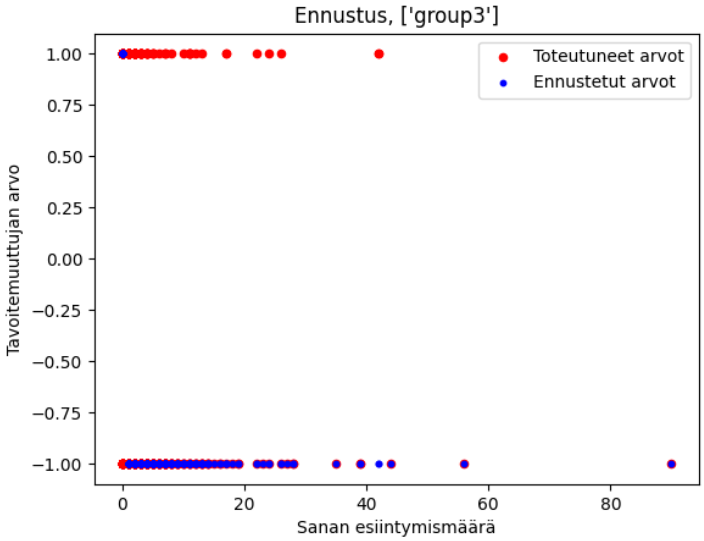
Ennustusten tulokset on esitetty kuvissa 35-42.



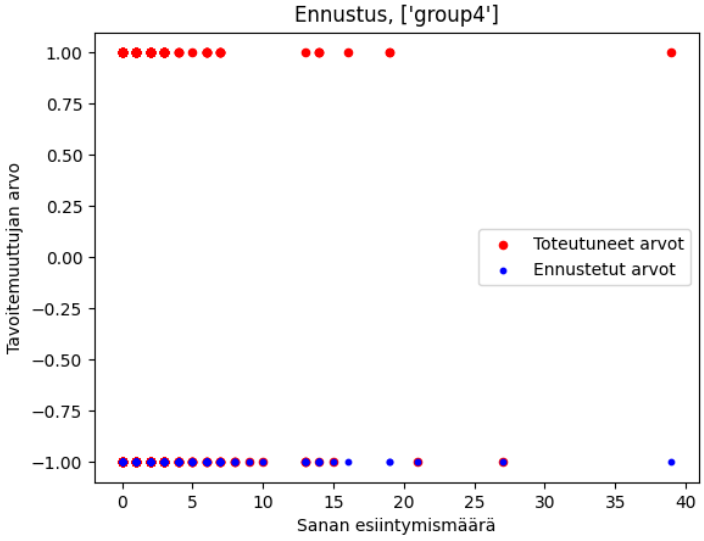
Kuva 35: Ennuste sanaryhmän 1 esiintymismäärän perusteella



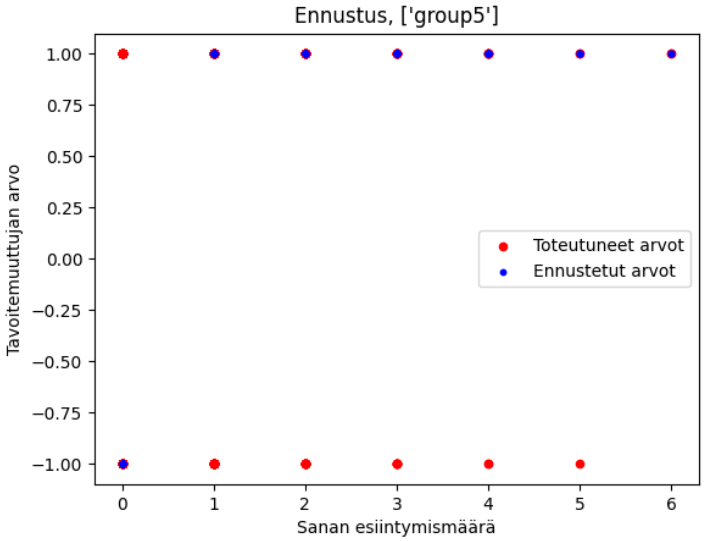
Kuva 36: Ennuste sanaryhmän 2 esiintymismäärän perusteella



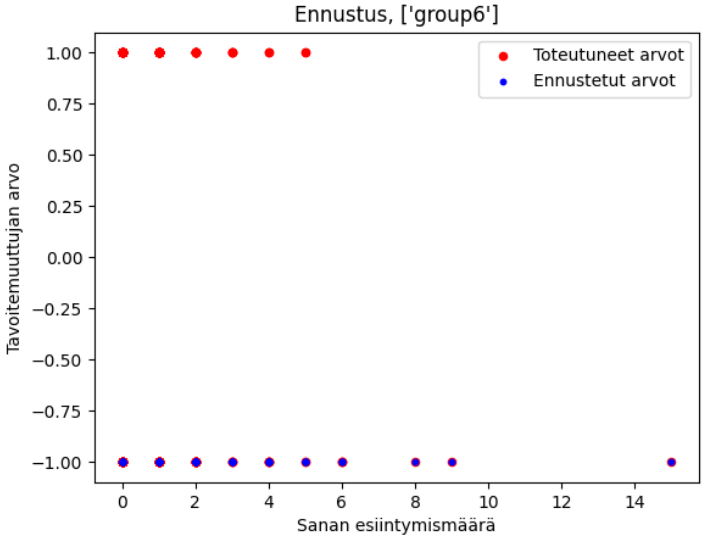
Kuva 37: Ennuste sanaryhmän 3 esiintymismäärän perusteella



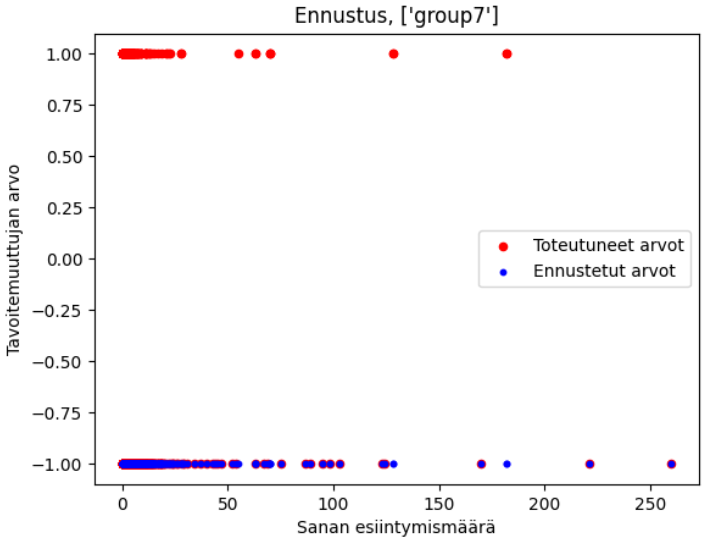
Kuva 38: Ennuste sanaryhmän 4 esiintymismäärän perusteella



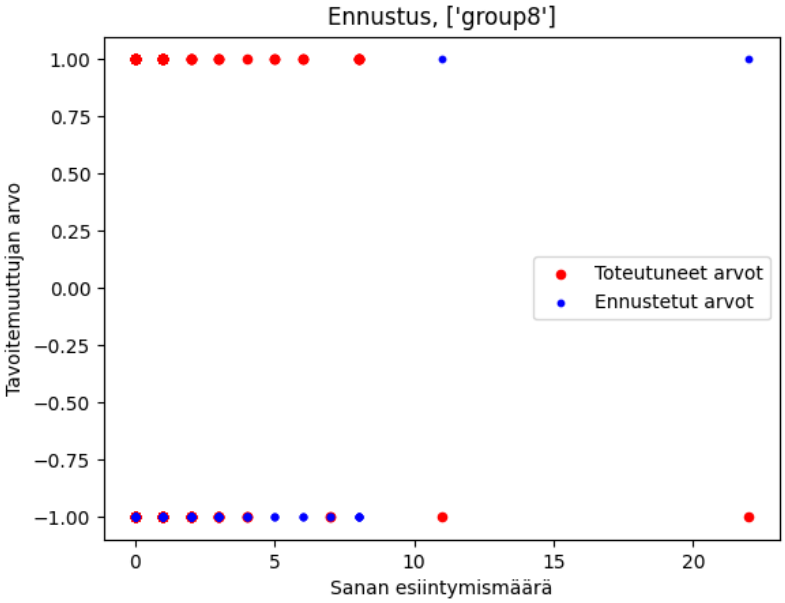
Kuva 39: Ennuste sanaryhmän 5 esiintymismäärän perusteella



Kuva 40: Ennuste sanaryhmän 6 esiintymismäärän perusteella



Kuva 41: Ennuste sanaryhmän 7 esiintymismäärän perusteella



Kuva 42: Ennuste sanaryhmän 8 esiintymismäärän perusteella

Kuvien perusteella parhaan ennustuksen tarjoaa ryhmä 2. Kuitenkin kaikkien versioiden keskimääräiset neliövirheet ovat melko suuria ottaen huomioon sen, että tavoitemuuttajan arvot ovat 1 tai -1. Niinpä kovin onnistuneita ennustuksia ei tällä tavoin saatu aikaiseksi.