Matemaattiset ja tilastolliset ohjelmistot, Kevät 2021

Maria Seppänen, maria.seppanen@tuni.fi

30.4.2021

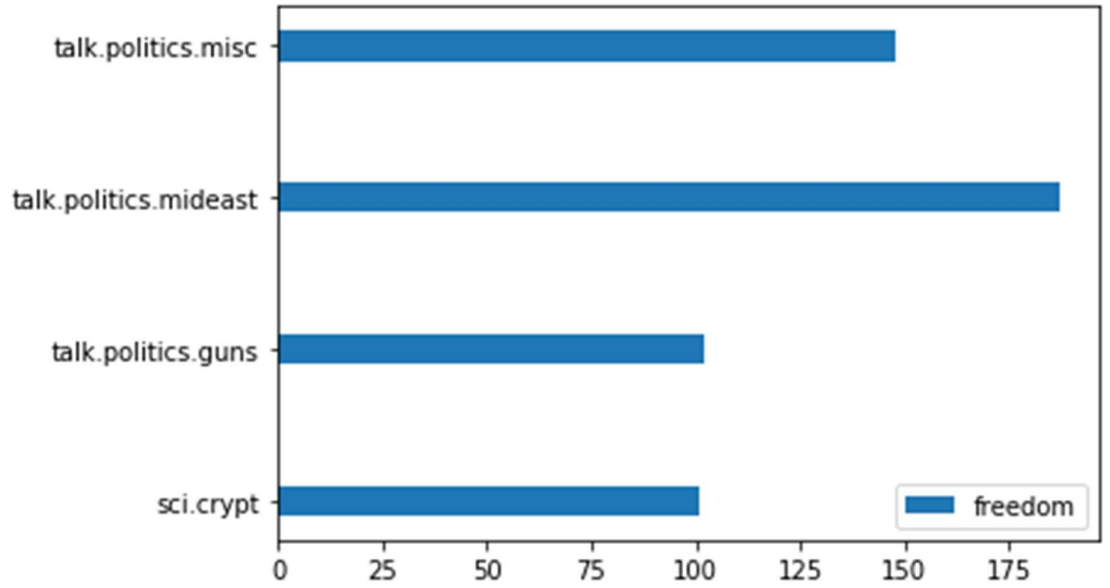
Python-moduuli: Harjoitustyö

Osa 1. Sanojen esiintymismäärät

Freedom:

- esiintyy eniten ryhmässä talk.politics.mideast ja vähiten ryhmässä sci.crypt

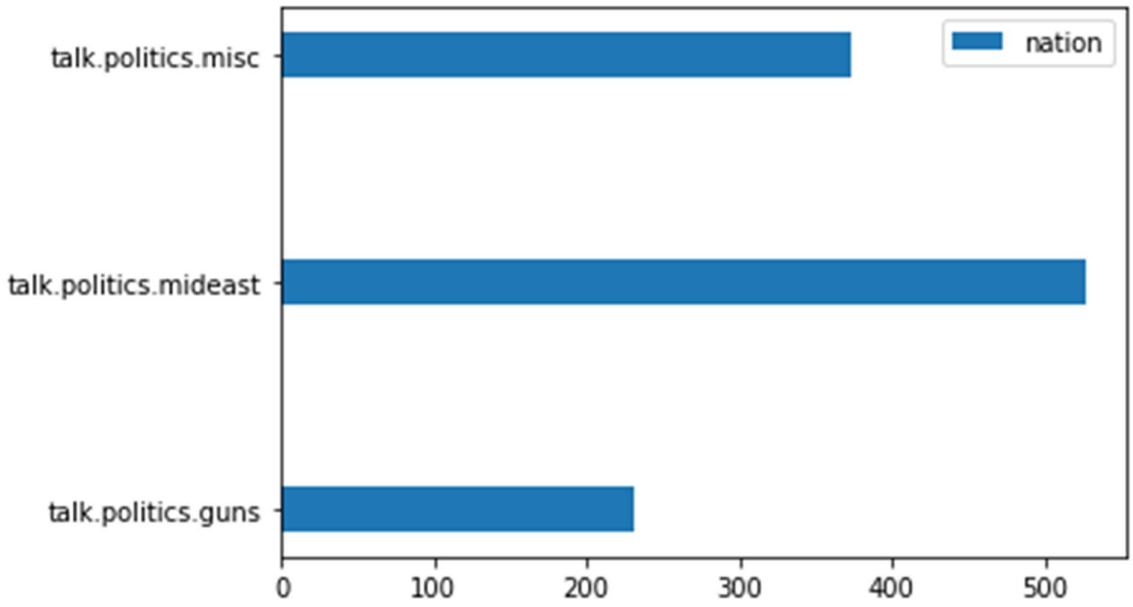
```
sci_crypt
freedom:
lukumäärä: 101.0
16.0
keskiarvo: 0.10369609856262833, mediaani: 0.0, keskihajonta: 0.6849780711152174
0,01% kvartiili: 0.0, 0,99% kvartiili: 7.243000000001416

talk.politics.guns
freedom:
lukumäärä: 102.0
10.0
keskiarvo: 0.10526315789473684, mediaani: 0.0, keskihajonta: 0.5652315749870317
0,01% kvartiili: 0.0, 0,99% kvartiili: 7.096000000000458

talk.politics.mideast
freedom:
lukumäärä: 187.0
10.0
keskiarvo: 0.19560669456066945, mediaani: 0.0, keskihajonta: 0.7660475482605262
0,01% kvartiili: 0.0, 0,99% kvartiili: 7.135000000000218

talk.politics.misc
freedom:
lukumäärä: 148.0
14.0
keskiarvo: 0.15132924335378323, mediaani: 0.0, keskihajonta: 0.8040349065784154
0,01% kvartiili: 0.0, 0,99% kvartiili: 10.092000000000553
```



Nation:

- esiintyy eniten ryhmässä talk.politics.mideast ja vähiten ryhmässä talk.politics.guns

```
talk.politics.guns
nation:
lukumäärä: 231.0
22.0
keskiarvo: 0.23839009287925697, mediaani: 0.0, keskihajonta: 0.9652708525303841
0,01% kvartiili: 0.0, 0,99% kvartiili: 7.480000000002292

talk.politics.mideast
nation:
lukumäärä: 527.0
17.0
keskiarvo: 0.551255230125523, mediaani: 0.0, keskihajonta: 1.5601344061188653
0,01% kvartiili: 0.0, 0,99% kvartiili: 17.0

talk.politics.misc
nation:
lukumäärä: 373.0
11.0
keskiarvo: 0.38139059304703476, mediaani: 0.0, keskihajonta: 1.076757951241296
0,01% kvartiili: 0.0, 0,99% kvartiili: 8.069000000000415
```



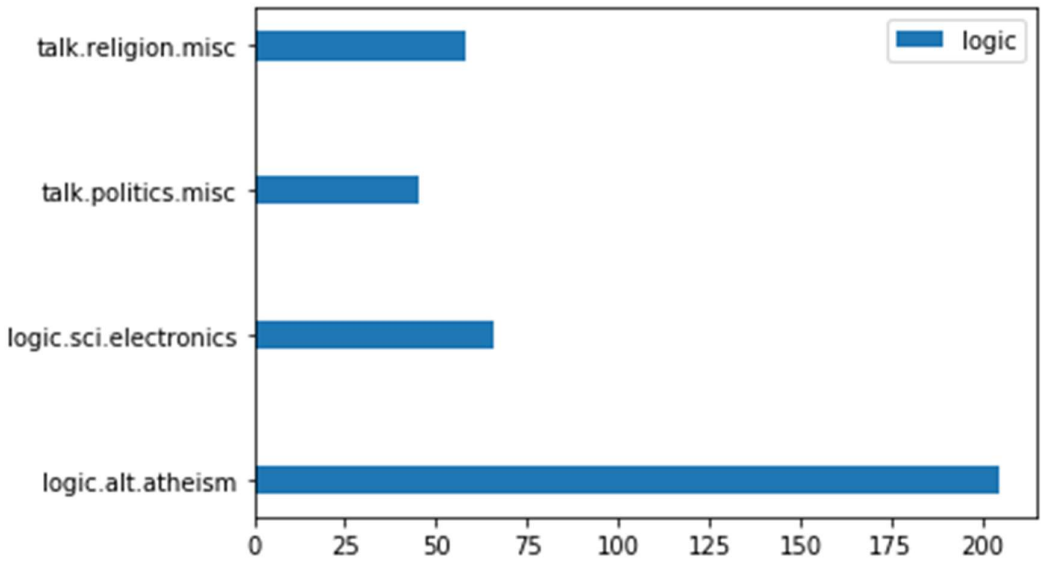
Maria Seppänen, maria.seppanen@tuni.fi

30.4.2021

Logic:

- esiintyy eniten ryhmässä logic.alt.atheism ja vähiten ryhmässä talk.politics.misc

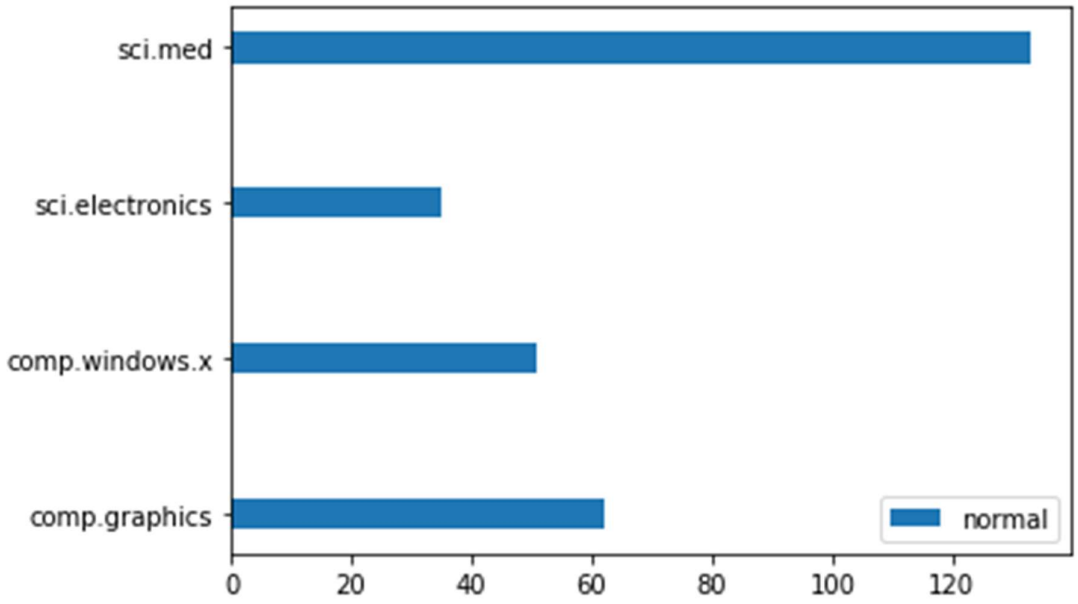
```
logic.alt.atheism
logic:
lukumäärä: 205.0
19.0
keskiarvo: 0.2104722792607803, mediaani: 0.0, keskihajonta: 0.9009833515357779
0,01% kvartiili: 0.0, 0,99% kvartiili: 8.29700000000173

logic.sci.electronics
logic:
lukumäärä: 66.0
10.0
keskiarvo: 0.072607260726072261, mediaani: 0.0, keskihajonta: 0.5604512821279437
0,01% kvartiili: 0.0, 0,99% kvartiili: 7.276000000000295

talk.politics.misc
logic:
lukumäärä: 45.0
4.0
keskiarvo: 0.046012269938650305, mediaani: 0.0, keskihajonta: 0.28423152846500294
0,01% kvartiili: 0.0, 0,99% kvartiili: 3.0230000000001382

logic.talk.religion
logic:
lukumäärä: 58.0
6.0
keskiarvo: 0.060353798126951096, mediaani: 0.0, keskihajonta: 0.36012075478904604
0,01% kvartiili: 0.0, 0,99% kvartiili: 5.040000000000077
```



Normal:

- esiintyy eniten ryhmässä sci.med ja vähiten ryhmässä sci.electronics

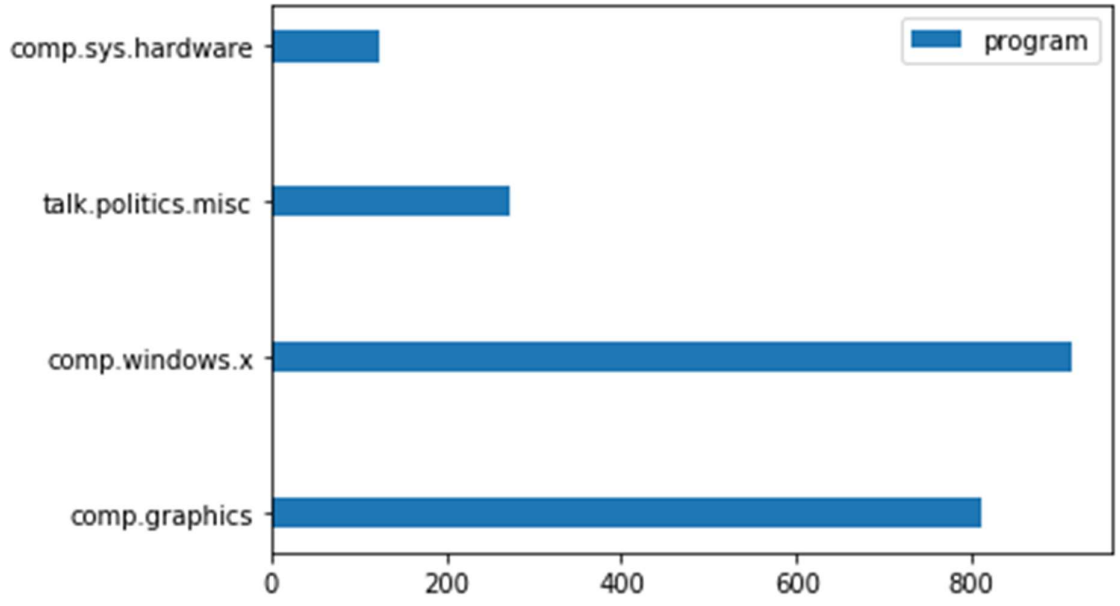
```
comp.graphics
normal:
lukumäärä: 62.0
9.0
keskiarvo: 0.06958473625140292, mediaani: 0.0, keskihajonta: 0.5124141433728697
0,01% kvartiili: 0.0, 0,99% kvartiili: 6.330000000000382

comp.windows.x
normal:
lukumäärä: 51.0
6.0
keskiarvo: 0.05622932745314223, mediaani: 0.0, keskihajonta: 0.3257396140120866
0,01% kvartiili: 0.0, 0,99% kvartiili: 3.282000000000153

sci.electronics
normal:
lukumäärä: 35.0
4.0
keskiarvo: 0.03850385038503851, mediaani: 0.0, keskihajonta: 0.2430812785258023
0,01% kvartiili: 0.0, 0,99% kvartiili: 3.0920000000000982

sci.med
normal:
lukumäärä: 133.0
12.0
keskiarvo: 0.1453551912568306, mediaani: 0.0, keskihajonta: 0.6546147910686233
0,01% kvartiili: 0.0, 0,99% kvartiili: 5.602000000000885
```



Program:

- esiintyy eniten ryhmässä com.windows.x ja vähiten ryhmässä comp.sys.hardware

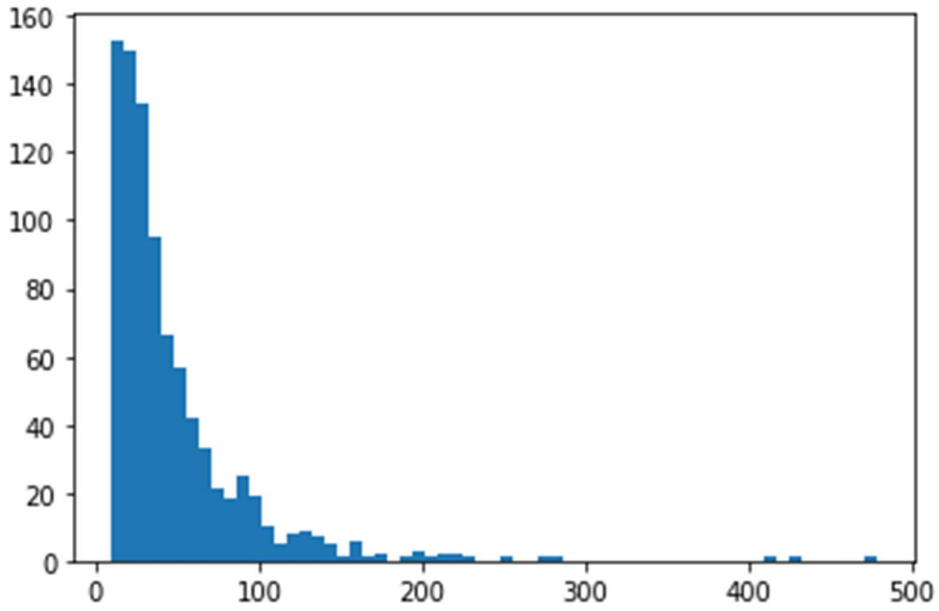
```
comp.graphics
program:
lukumäärä: 812.0
47.0
keskiarvo: 0.9113355780022446, mediaani: 0.0, keskihajonta: 3.8397675945972534
0,01% kvartiili: 0.0, 0,99% kvartiili: 47.0

comp.windows.x
program:
lukumäärä: 914.0
78.0
keskiarvo: 1.007717750826902, mediaani: 0.0, keskihajonta: 5.223371021642908
0,01% kvartiili: 0.0, 0,99% kvartiili: 78.0

talk.politics.misc
program:
lukumäärä: 272.0
51.0
keskiarvo: 0.278118609406953, mediaani: 0.0, keskihajonta: 2.080358931475062
0,01% kvartiili: 0.0, 0,99% kvartiili: 30.483000000002903

comp.sys.hardware
program:
lukumäärä: 123.0
34.0
keskiarvo: 0.13369565217391305, mediaani: 0.0, keskihajonta: 1.3022770025972132
0,01% kvartiili: 0.0, 0,99% kvartiili: 19.296000000002095
```



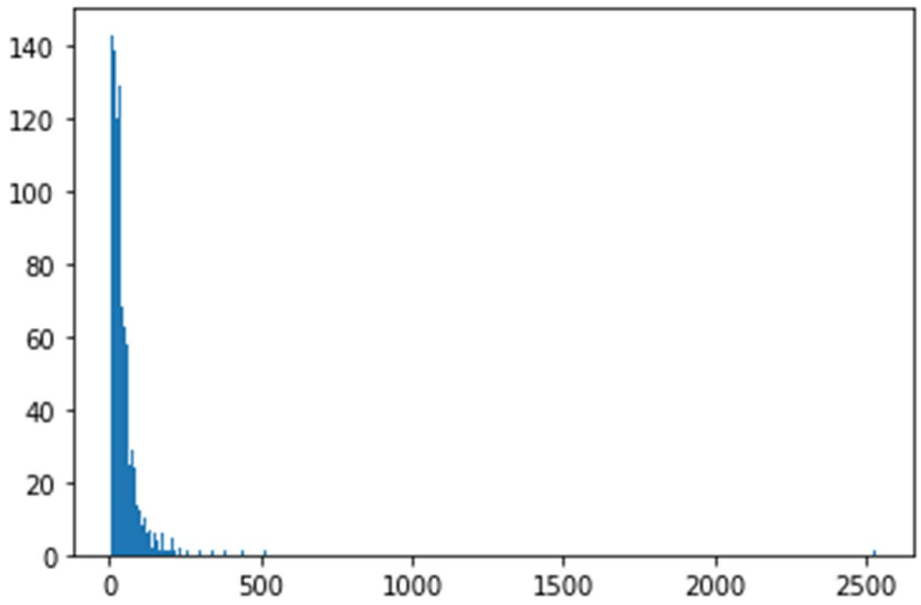
Osa 2: Viestien pituudet

Histogrammit

Histogrammi viestien pituuksista ryhmissä rec.sport.baseball:

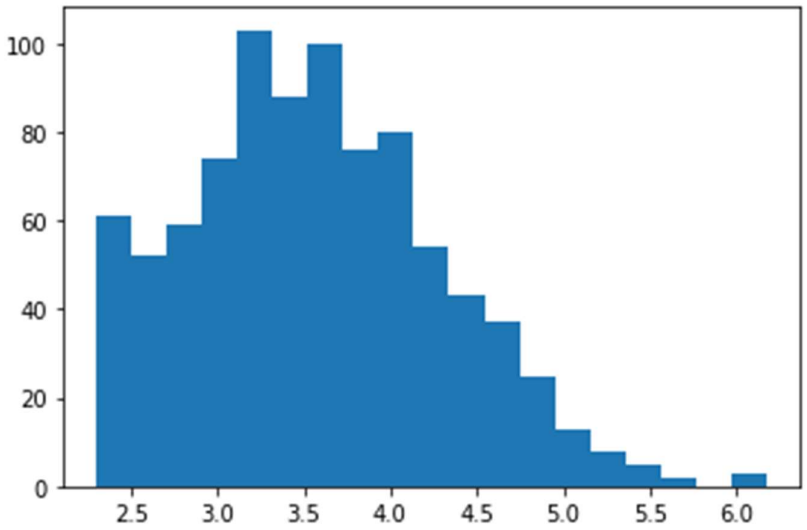


Histogrammi viestien pituuksista ryhmissä rec.sport.hockey:

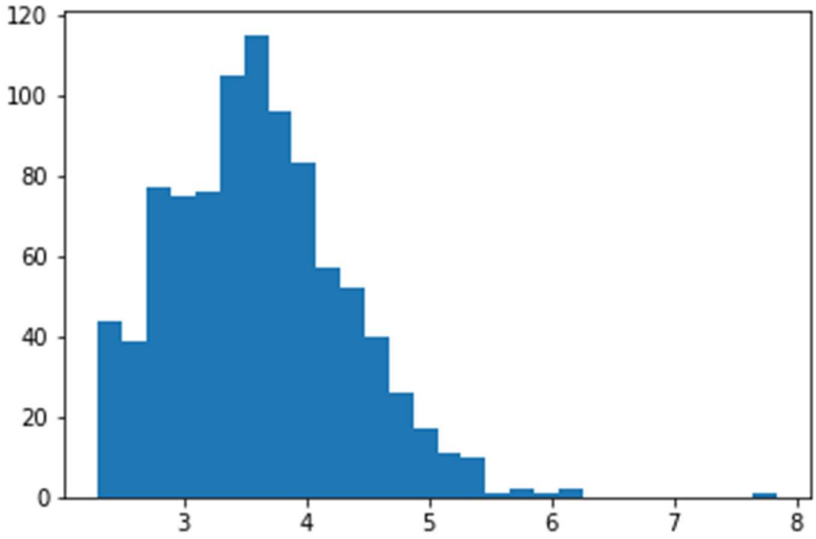


Logaritmiset histogrammit

Histogrammi mutta pituuksien logaritmeista ryhmässä rec.sport.baseball:



Histogrammi mutta pituuksien logaritmeista ryhmässä rec.sport.hockey:



Logaritmiset kuvaajat näyttävät olevan lähempänä normaalijakautunutta.

Tilastolliset t-testit

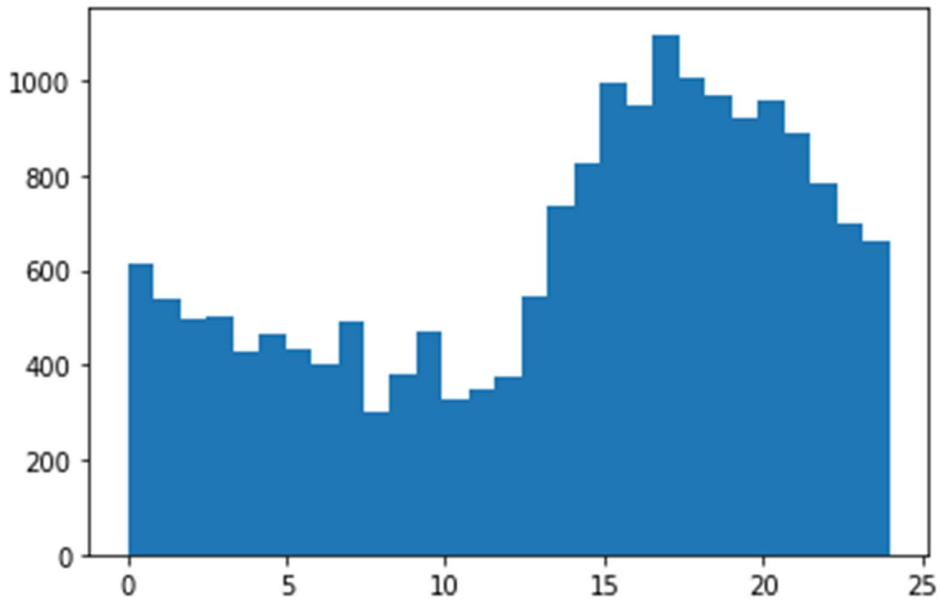
Viestien logaritmisten pituuksien odotusarvojen t-testiksi ryhmien rec.sport.baseball ja rec.sport.hockey välillä tuli p-arvoksi 0.24743033891447788, joka on suurempi kuin merkitsevyystaso 0.05. Täten nollahypoteesi jää voimaan, eli logaritmisten pituuksien jakaumat eivät eroa 5% merkitsevyystasolla.

Saman testin, mutta ryhmien rec.autos ja rec.motorcycles välillä tuli p-arvoksi tuli 6.070535775703711e-05, joka on pienempi kuin merkitsevyystaso 0.05. Täten nollahypoteesi hylätään ja vaihtoehtoinen hypoteesi astuu voimaan, eli logaritmisten pituuksien jakaumat eroavat toisistaan 5%-merkitsevyystasolla.

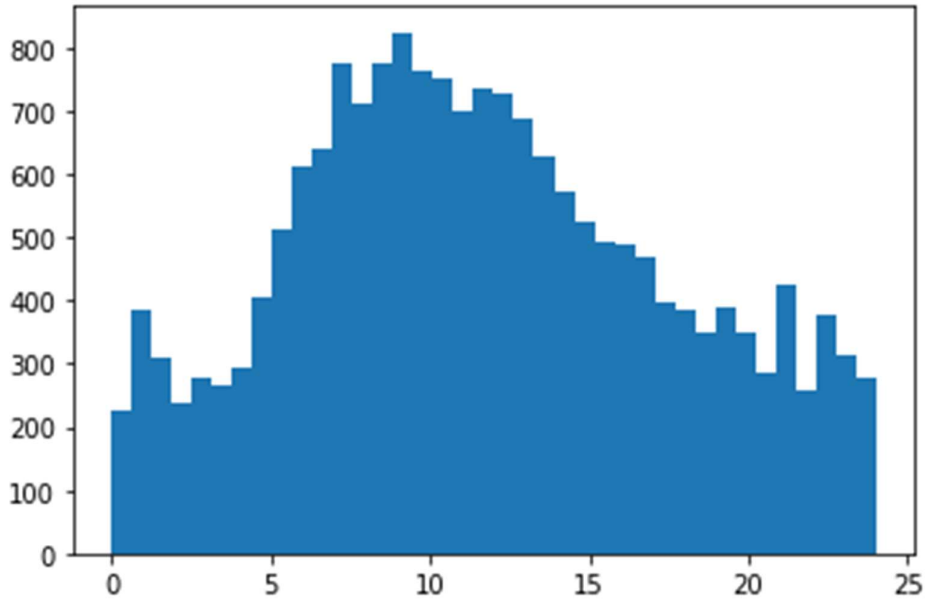
Osa 3: Kirjoitusajat

Histogrammit kirjoitusajoista

Viestien kirjoitusajat tunteina keskiyöstä:

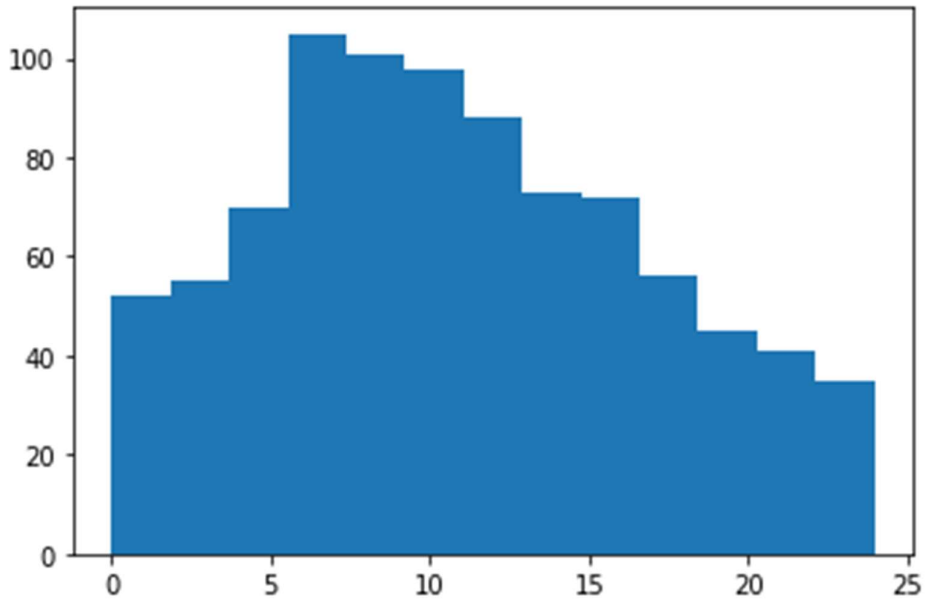


Viestien kirjoitusajat tunteina kahdeksasta aamulla:

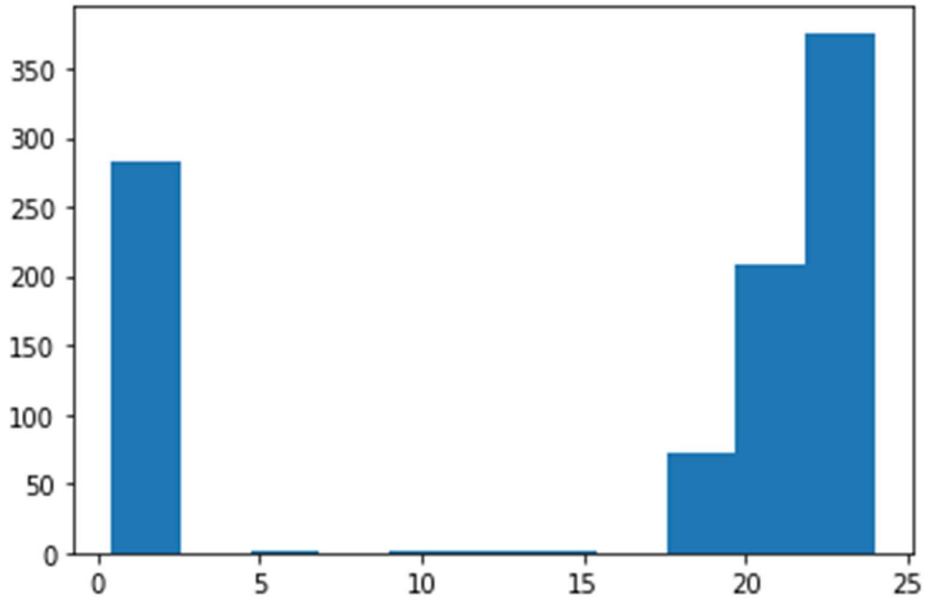


Kirjoitusajan keskiarvo: 11.67807530519459, mediaani: 11.202083333333334 ja keskihajonta: 5.885242031525389. Jakauma näyttää enemmän normaalijakautuneelta.

Kirjoitusajat ryhmässä comp.graphics:



Kirjoitusajat ryhmässä soc.religion.christian:



Comp.graphics keskiarvo: 10.967666791370508

Soc.religion.christian keskiarvo: 15.56237872915196

Kirjoitusajat vaikuttavat hyvin erilaisilta, comp.graphis on aktiivinen aika lailla ympäri vuorokauden, kun taas soc.religion.christianissa viestejä kirjoitetaan lähinnä 17-03 aikavälillä ja hyvin vähän muulloin.

T-testi

Keskimääräisten kirjoitusaikojen eroa testasin t-testillä, jonka p-arvoksi tuli 1.5025984582022796e-33, jolloin nollahypoteesi hylätään, eli kirjoitusajoissa on merkitsevä ero.

Osa 4: Korrelaatiot

Sanojen jpeg ja gif korrelaatio: 0.9659556335564533.

Sanojen write ja sale korrelaatio: -0.081973210676033.

Sanojen jpeg ja gif korrelaatio ryhmässä comp.graphics: 0.9913782838589174, mikä on suurempi kuin korrelaatio kaikkiaan.

Osa 5: Sentimentin analysointi

Sentimentin normaalisuuden testi

Sentimenttiarvon normaalisuuden testin p-arvo on niin pieni, että se pyöristyy nollaan. Tällöin nollahypoteesi hylätään ja sentimentti ei ole normaalijakautunut.

Sentimentin jakauma

Ryhmän 1 keskiarvo: 0.0086, mediaani: 0.0102, keskihajonta: 0.0725,

25% kvartiili: -0.0337 ja 75% kvartiili 0.0530.

Ryhmän 2 keskiarvo: 0.0387, mediaani: 0.0357, keskihajonta: 0.0573,

25% kvartiili: 0.0000 ja 75% kvartiili 0.0690.

Ryhmän 3 keskiarvo: 0.0254, mediaani: 0.0309, keskihajonta: 0.0822,

25% kvartiili: 0.0000 ja 75% kvartiili 0.0606.

Ryhmän 4 keskiarvo: 0.0216, mediaani: 0.0207, keskihajonta: 0.0594,

25% kvartiili: -0.0085 ja 75% kvartiili 0.0541.

Ryhmän 5 keskiarvo: 0.0308, mediaani: 0.0299, keskihajonta: 0.0645,

25% kvartiili: 0.0000 ja 75% kvartiili 0.0676.

Ryhmän 6 keskiarvo: 0.0276, mediaani: 0.0246, keskihajonta: 0.0586,

25% kvartiili: 0.0000 ja 75% kvartiili 0.0593.

Ryhmän 7 keskiarvo: 0.0400, mediaani: 0.0382, keskihajonta: 0.0537,

25% kvartiili: 0.0000 ja 75% kvartiili 0.0724.

Ryhmän 8 keskiarvo: 0.0096, mediaani: 0.0152, keskihajonta: 0.0593,

25% kvartiili: -0.0257 ja 75% kvartiili 0.0468.

Ryhmän 9 keskiarvo: 0.0013, mediaani: 0.0000, keskihajonta: 0.0592,

25% kvartiili: -0.0339 ja 75% kvartiili 0.0370.

Ryhmän 10 keskiarvo: 0.0262, mediaani: 0.0270, keskihajonta: 0.0605,

25% kvartiili: -0.0023 ja 75% kvartiili 0.0608.

Ryhmän 11 keskiarvo: 0.0259, mediaani: 0.0234, keskihajonta: 0.0582,

25% kvartiili: -0.0065 ja 75% kvartiili 0.0592.

Ryhmän 12 keskiarvo: 0.0170, mediaani: 0.0157, keskihajonta: 0.0598,

25% kvartiili: -0.0196 ja 75% kvartiili 0.0550.

Ryhmän 13 keskiarvo: 0.0357, mediaani: 0.0333, keskihajonta: 0.0551,

25% kvartiili: 0.0000 ja 75% kvartiili 0.0722.

Ryhmän 14 keskiarvo: -0.0045, mediaani: 0.0000, keskihajonta: 0.0715,

25% kvartiili: -0.0465 ja 75% kvartiili 0.0401.

Ryhmän 15 keskiarvo: 0.0104, mediaani: 0.0118, keskihajonta: 0.0557,

25% kvartiili: -0.0254 ja 75% kvartiili 0.0431.

Ryhmän 16 keskiarvo: 0.0237, mediaani: 0.0278, keskihajonta: 0.0779,

25% kvartiili: -0.0217 ja 75% kvartiili 0.0659.

Ryhmän 17 keskiarvo: -0.0204, mediaani: -0.0196, keskihajonta: 0.0658,

25% kvartiili: -0.0625 ja 75% kvartiili 0.0206.

Ryhmän 18 keskiarvo: -0.0217, mediaani: -0.0216, keskihajonta: 0.0645,

25% kvartiili: -0.0607 ja 75% kvartiili 0.0157.

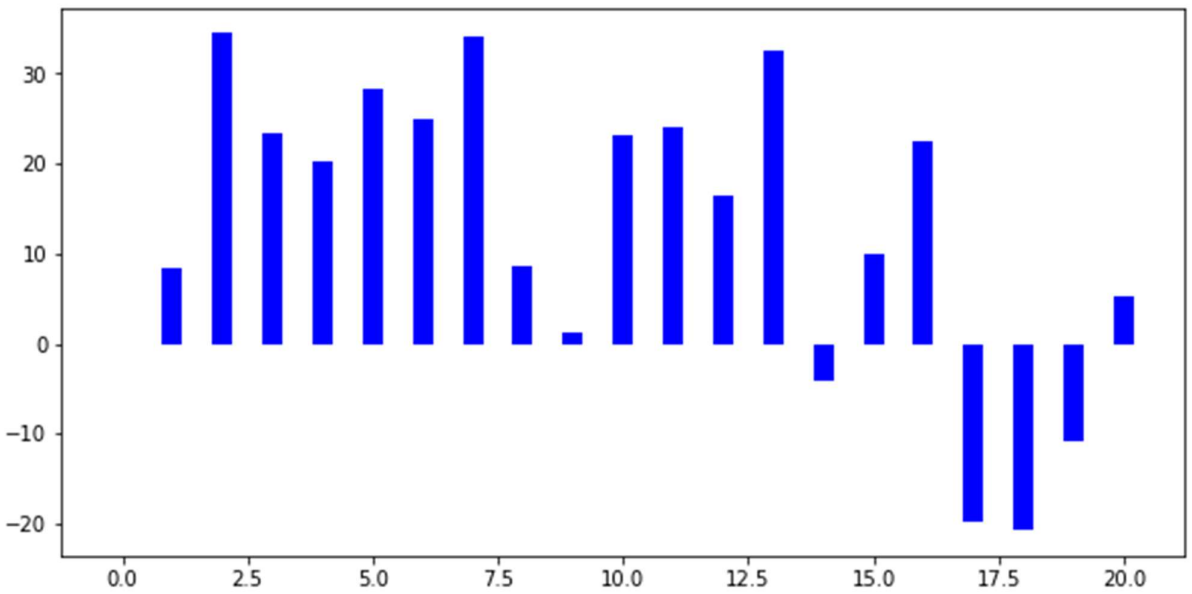
Ryhmän 19 keskiarvo: -0.0111, mediaani: -0.0097, keskihajonta: 0.0640,

25% kvartiili: -0.0509 ja 75% kvartiili 0.0286.

Ryhmän 20 keskiarvo: 0.0055, mediaani: 0.0041, keskihajonta: 0.0698,

25% kvartiili: -0.0370 ja 75% kvartiili 0.0496.

Histogrammi



T-testi

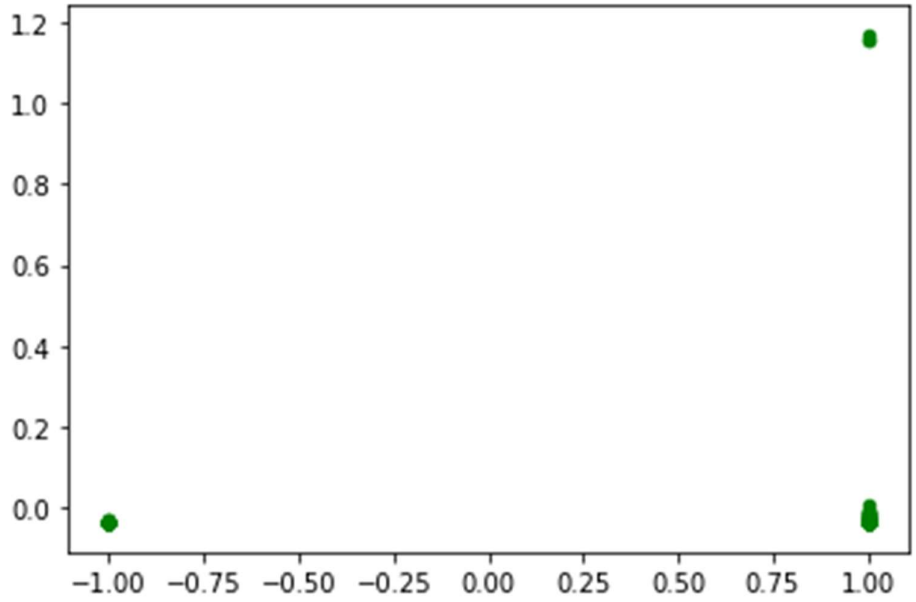
P-arvot testeittäin:

1. comp.sys.ibm.pc.hardware vs. comp.sys.hardware: 0.0014340022828717297.
2. rec.sport.baseball vs. rec.sport.hockey: 0.8983835913671558.
3. rec.autos vs. rec.motorcycles: 0.00273396989329422.

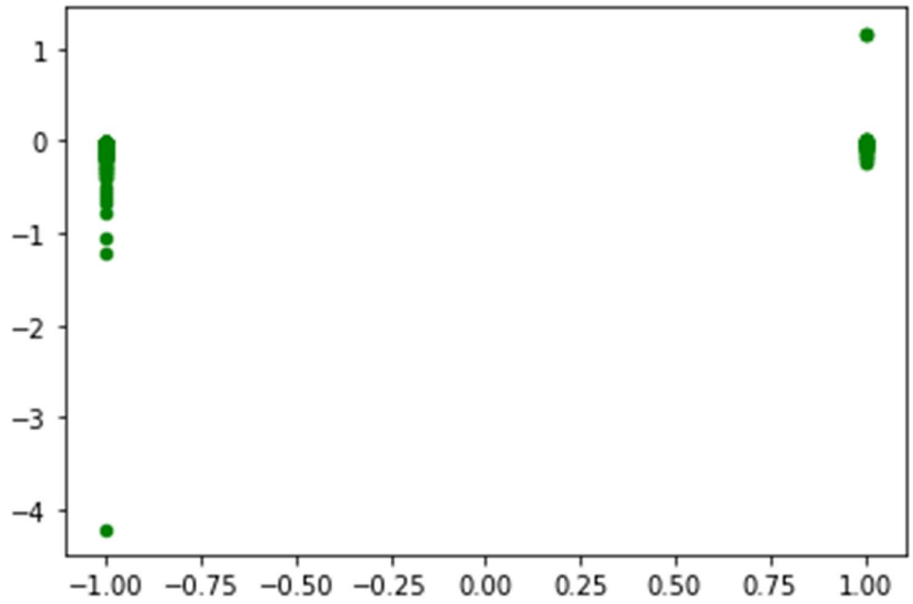
Kaikki testit, joissa p-arvo on pienempi kuin merkitsevyystaso 0.05, nollahypoteesi hylätään. Eli sentimentin jakauma poikkeaa testien 1 ja 3 ryhmien välillä tilastollisesti merkitsevästi. Tetsin 2 ryhmien sentimentin jakauma ei poikkea tilastollisesti merkitsevästi 5%-merkitsevyystasolla.

Osa 6: Uutisryhmän ennustaminen

a) Keskimääräinen neliövirhe: 0.996308174380203.



b) Keskimääräinen neliövirhe: 0.9832151724253745.



c) Painokertoimet ovat: -0.05598364, 0.01628368, -0.03494182, -0.01823374, 0.08612578, 0.10117179, -0.01780627, -0.01186285. Eniten vaikuttavat sanaryhmät 6 ja 5. Keskimääräinen neliövirhe: 0.9084343163510733. Neliövirhe on pienempi kuin kahdessa edellisessä mallissa, joten se on parempi kuin mallit kohdissa a) ja b).