Elyas Addawe – Saku Ihalainen – Hanna Kaimo

Kirjojen klusterointi (CRISP-DM)

Tehtävä n:o 10

Metropolia Ammattikorkeakoulu
Tieto- ja viestintätekniikka
Datan käsittely ja koneoppiminen
25.4.2022

1 Tavoitteet

Pyrimme selvittämään voidaanko tekstinlouhinnan tekniikoin saada selville, kuka/ kumpi kirjailija on kirjoittanut jonkin teoksen.

2 Data

Data on kirjailijan kirjoittamaa tekstiä. Tehtävänä oli hakea kolme teosta kahdelta kirjailijalta. Latasimme 6 eri teosta sivusta https://www.gutenberg.org/.

Ensimmäinen kirjailija on **Mary Wollstonecraft (Godwin) Shelley**

Valitsimme seuravaat kolme teosta häneltä:

"Frankenstein; or, the Modern Prometheus"

"The last man"

"Tales and stories"

Toisen kirjailijan nimi on **Lewis Carrol**

Valitsin seuraavat kolme teosta häneltä:

Alice Adventures in Wonderland

Alice Adventures underground

Through The Looking glass

Datan muoto on Nominaali.

Kirjoja pitää yhdistää ja poistaa erikoismerkit wekan käyttöä varten

Seuraavaksi muodostettiin yhtenäinen tiedosto aineisto kuudesta eri kirjasta.

3 Datan valmistelu

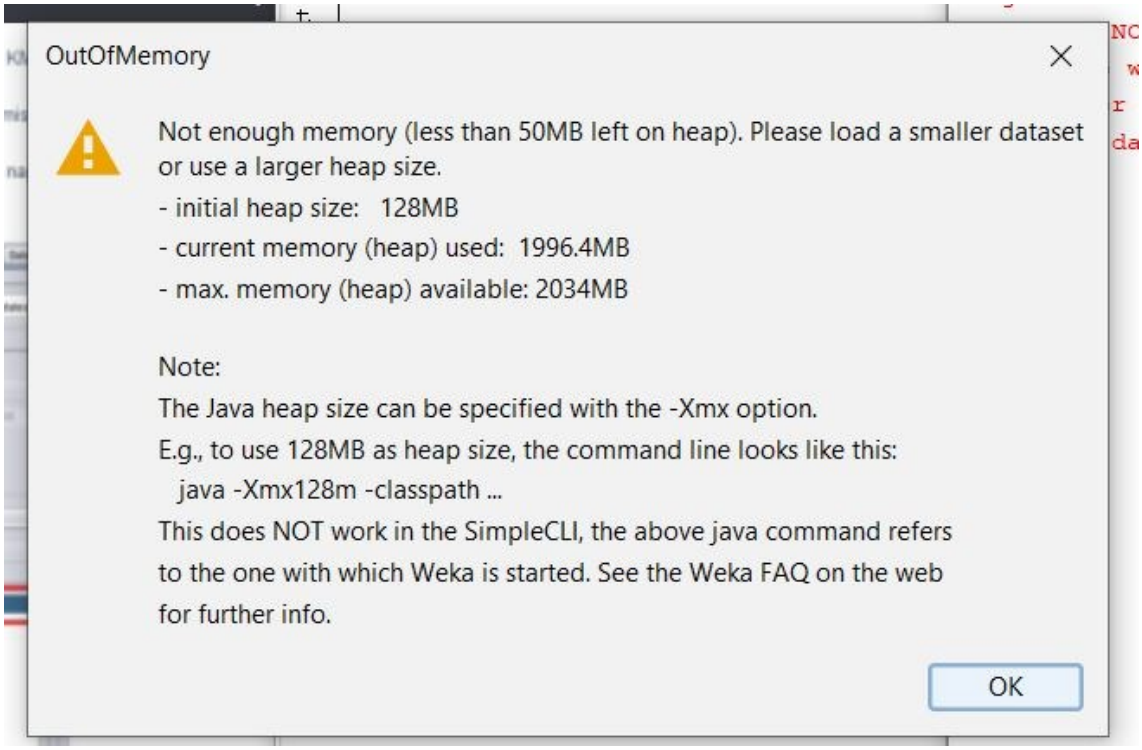
Teimme datalle niin paljon valmisteluja, että emme muista niitä kaikkia nyt, mutta laitamme tässä mitä muistamme, koska emme tienneet mitä kautta asia alkaa toimimaan. Ensimmäisenä muutimme csv- muodossa sivulta otetut datat (kaksi erillistä tiedostos, yksi kummaltakin kirjailijalta) väärästä csv- muodosta teksti-tiedostomuotoon. Poistimme turhat tyhjät rivit notepadissä (edit- line operatios-remove empty lines). Teimme muutokset, joiden jälkeen saimme konvertoitua datan oikeaan csv- muotoon. Putsasimme dataa, poistimme turhia merkkejä (esim. turhia hipsuja, erikoiserkkejä) jne. molemmista datoista. Yhdistimme datat, siirsimme ne samaan csv-tiedostoon, jolle loinmmekuvaavahkon nimen. Teimme datasta sellaisen, että saimme sen Wekassa muutettua arff- muotoon. Sen jälkeen aloitimme datan siistimisen Wekassa. Wekassa teimme datalle ne toimenpiteet, jotka oli suosituksena kyseisellä tunnila annettu ja katsoimme muuten neuvoa YouTubesta (esimerkkeinä NominalToString, StringToWordVector).Wekan avulla myös laitoimme kirjaimet pieniksi.

4 Mallinnus

Analysoimme dataa Wekan k-means Algoritmilla (Simple KMeans). Algoritmi löytyi klustereista. Cluster Moden arvoksi laitoimme "Use Training Set". Seediksi laitoimme attribuuttien määrän: 10 ja NumClusteriksi: 2. Tuloksia tuli 2 kappaletta. Tulokseen tuli mukaan kaikki attribuutit.

Laitoimme kaksi clusteria sillä tiedostoaineisto koostui kahdesta eri kirjailijasta.

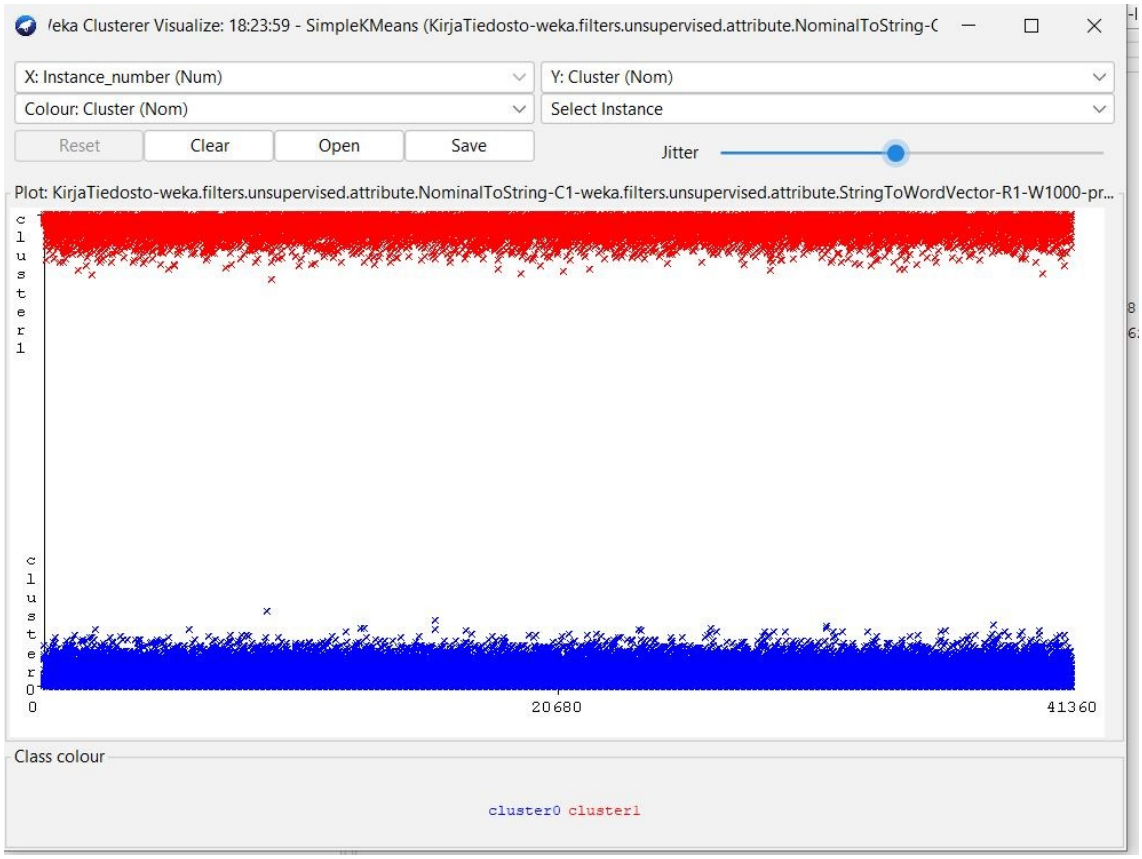
Huomasimme että weka kaatui koska muistia ei riittänyt. Vähensimme max-iteraattoria 250 alunperin 500;


OutOfMemory
Not enough memory (less than 50MB left on heap). Please load a smaller dataset or use a larger heap size.
- initial heap size: 128MB
- current memory (heap) used: 1996.4MB
- max. memory (heap) available: 2034MB
Note:
The Java heap size can be specified with the -Xmx option.
E.g., to use 128MB as heap size, the command line looks like this:
java -Xmx128m -classpath ...
This does NOT work in the SimpleCLI, the above java command refers to the one with which Weka is started. See the Weka FAQ on the web for further info.
OK

Tulokset:

```
kMeans
======

Number of iterations: 9
Within cluster sum of squared errors: 296180.6746051287

Initial starting points (random):

Cluster 0: {65 1.968981,121 1.863979,175 3.881386,531 3.973296,551 1.937558,617 0.75310
Cluster 1: {37 0.716772,59 1.82723,406 1.687647,450 3.846058,617 0.753105,625 2.823407,

Missing values globally replaced with mean/mode

Final cluster centroids:
                            Cluster#
Attribute          Full Data         0         1
                   (41361.0) (29871.0) (11490.0)
================================================
able                  0.0061    0.0002    0.0216
abode                 0.0092    0.0086    0.0107
about                 0.0386    0.0316    0.0571
above                 0.0116    0.0138    0.0061
absence               0.0083     0.009    0.0064
account               0.0083     0.007    0.0117
accustomed            0.0056    0.0031    0.0119
across                0.0076    0.0081    0.0061
act                   0.0082    0.0065    0.0126
added                 0.0122    0.0083    0.0224
admiration            0.0056    0.0059    0.0049
adrian                0.0236    0.0208     0.031
advanced              0.0063    0.0063    0.0064
affection             0.0132    0.0132    0.0131
```

```
Time taken to build model (full training da

=== Model and evaluation on training set ==

Clustered Instances

0       29871 ( 72%)
1       11490 ( 28%)
```



5 Arviointi

Kun seed oli asetettu 10:ksi clusterit jakautuivat 0 joka oli 72% ja 1 28%;

Voimme päätellä mikä kirjailija klusteri kuuluu mihinkin vertaamalla sanoja mitä käytetään tekstissä. Klusteri 0 taitaa olla Mary Wollstonecraft (Godwin) Shelley sillä hänellä on tiettyjä sanoja mitkä esiintyy vain hänen kirjoissaan. Klusteri 1 taas on Lewis Carroll.

Odotimme että klusterit jakautuisivat tasaisemmin enemmän.

Suuri ero %voi johtua siitä että kirjailijan A teokset ovat paljon pidempiä kun Kirjailija B:n. Mary:n teokset olivat paljon pidempiä noin 70% yhteisestä aineistoista.

Jos klusteri tunnistaminen menee oikein niin. Tuntemattoman teoksen tunnistaminen tapahtuu lisäämällä uusi teos aineistoon ja seuraamalla kumpi klusteri kasvaa enemmän.

6 Täytäntöönpano

Mitä tulokset tarkoittavat tavoitteiden näkökulmasta? Miten vastaatte ensimmäisen kohdan kysymyksiin? Mikä on toimenpidesuositus, jos sellaisen voi laatia?

Selvitimme tulosteiden avulla mihin klusteriin kirjailija kuuluu. Samalla päädyimme että kirjailijaa voidaan ennustaa seuraamalla hänen sanojaan ja miten useasti hän käyttää niitä. Ongelma voi tuolla kun kahden eri kirjailijan teokset ovat samanlaisia tai saman kirjailijan teokset eroavat täysin. Esimerkiksi kun kirjailija kirjoittaa yksittäisen kirjan eri kielellä kuin mitä normaalisti kirjoittaa . Samalla kirjailija voi kirjoittaa kokonaan eri aiheesta joten sanat eivät ole samanlaisia kuin muissa teoksissa. Tämän tutkimuksen perusteella tiedonlouhinta tekniikan avulla saadaan irti tällaisesta laajasta datasta avain tietoja.