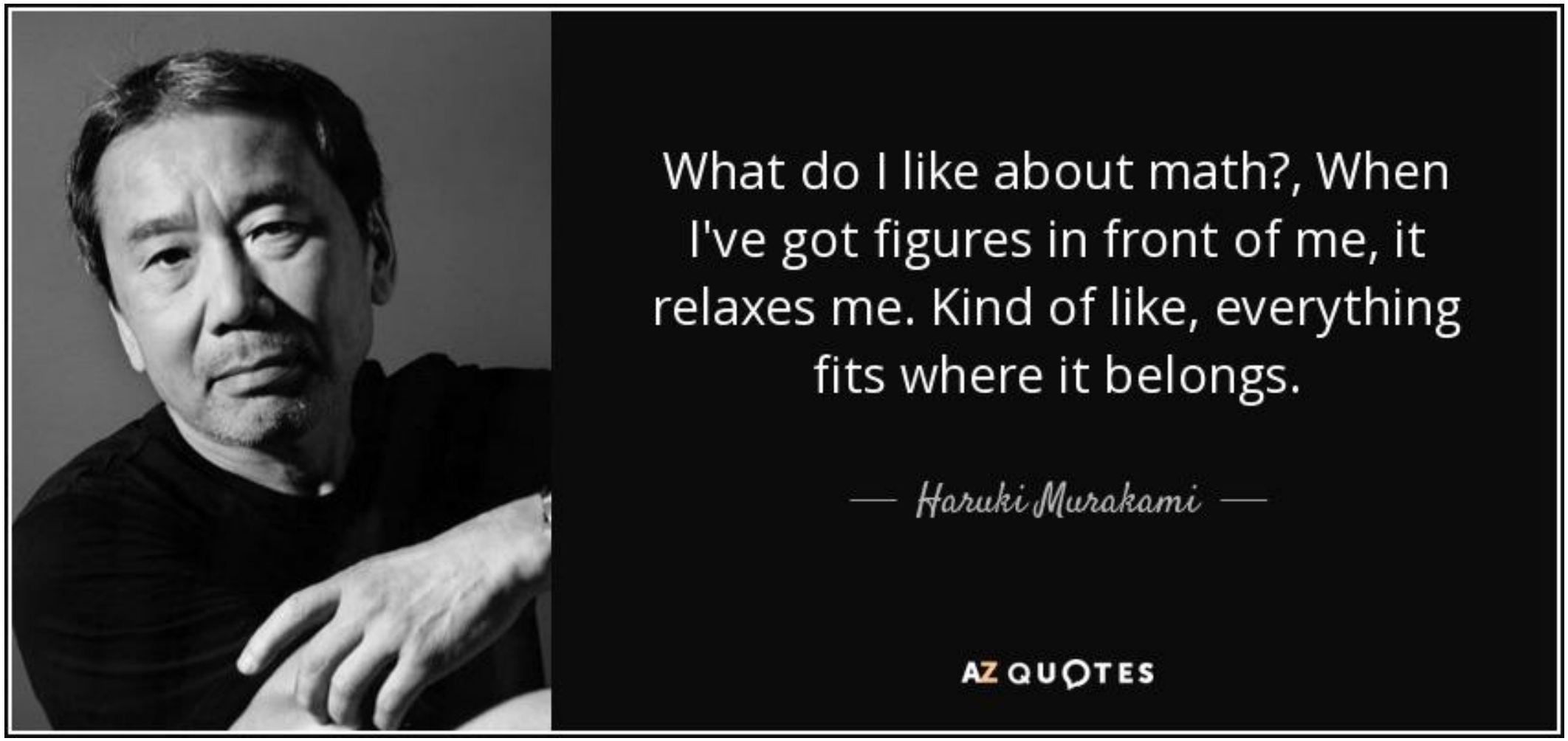
Пајтонски јуриш низ „Норвешка шума“

ШЕСТИ СЕМИНАР „МАТЕМАТИКА И ПРИМЕНИ“, 17 МАРТ 2023
ИНСТИТУТ ЗА МАТЕМАТИКА, ПРИРОДНО-МАТЕМАТИЧКИ ФАКУЛТЕТ, УНИВЕРЗИТЕТ „СВ. КИРИЛ И МЕТОДИЈ“, СКОПЈЕ

Марко Димовски, ЕВН Македонија

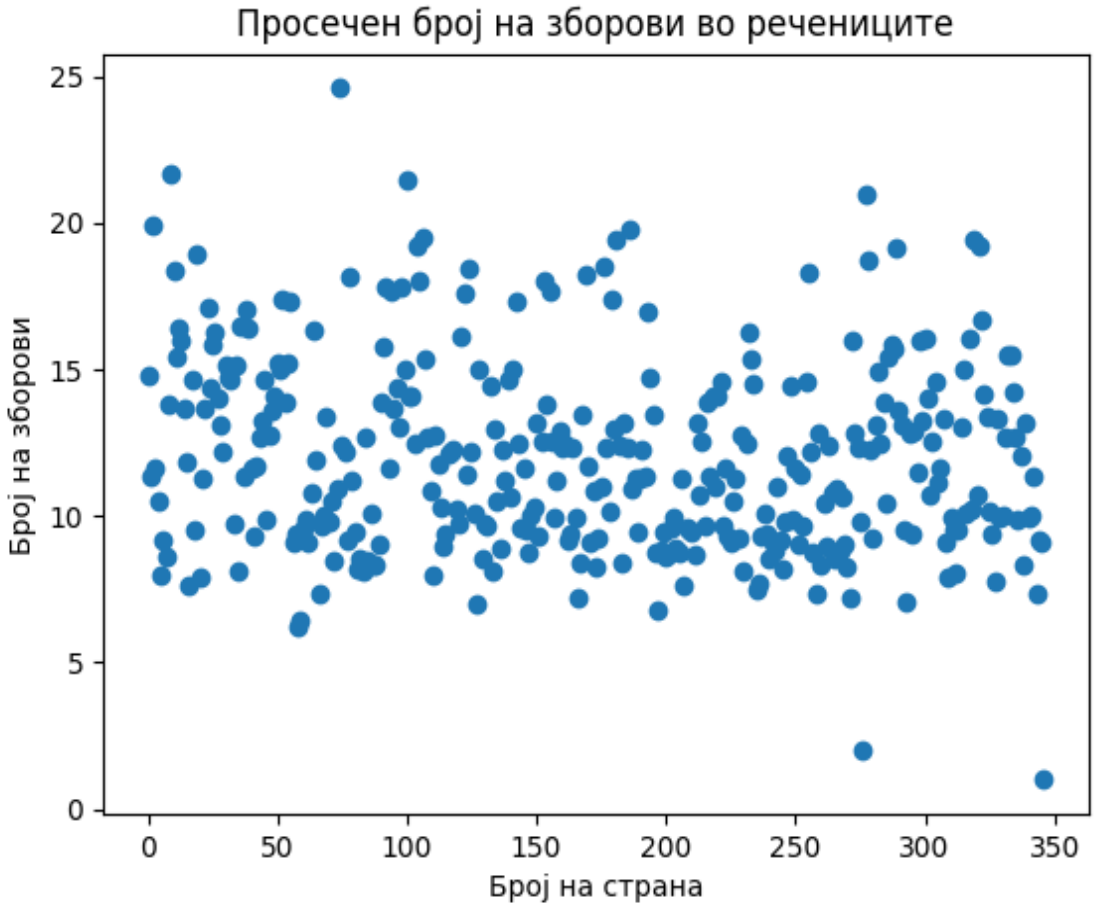

What do I like about math?, When I've got figures in front of me, it relaxes me. Kind of like, everything fits where it belongs.
— Haruki Murakami —
AZ QUOTES

Број на зборови во реченициите од книгата

- Просечниот број на зборови во реченициите од „Норвешка шума“ е околу 12 зборови во една реченица (просекот е 11,73 зборови).

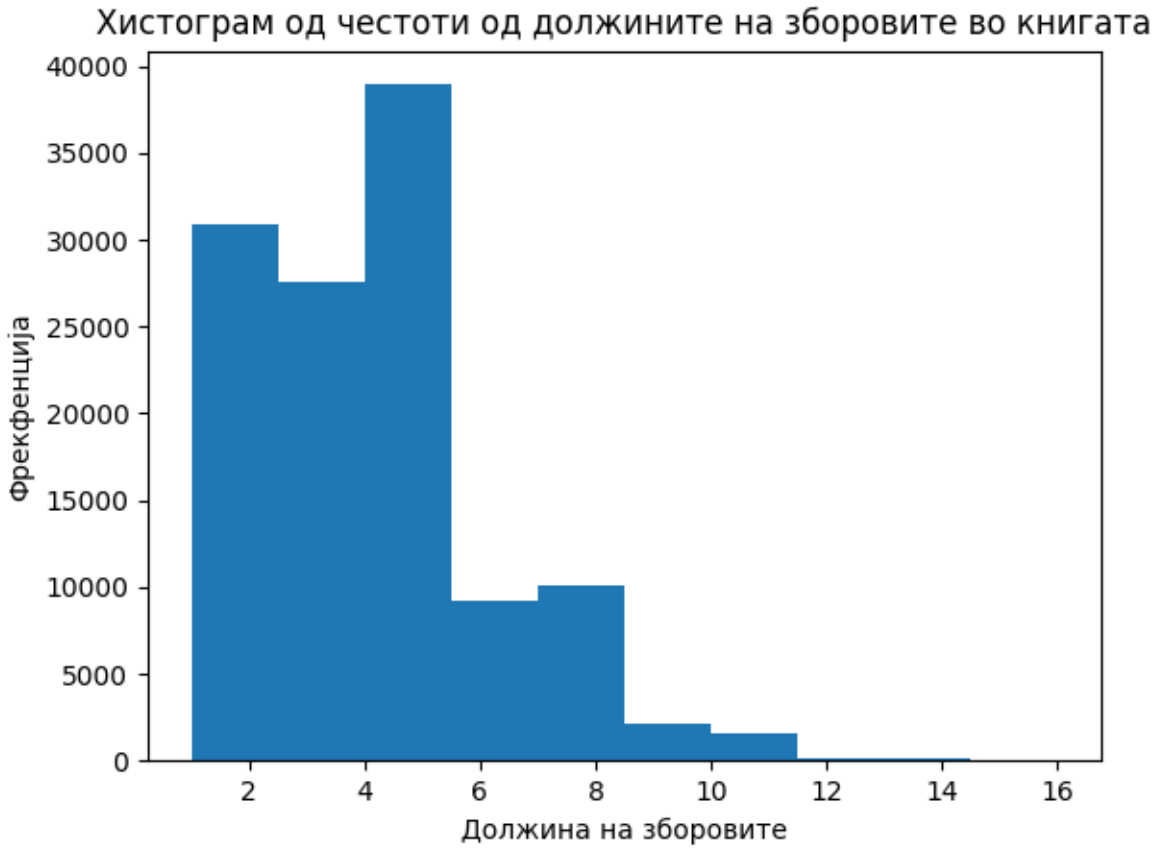
- Просек од 12 зборови во реченица имал и Шекспир во „Хамлет“ и „Макбет“. Во „Алиса во земјата на чудата“ реченициите имале просечна должина од 20 зборови, а во „Моби Дик“ од 26 зборови. Во „Загубениот рај“ од Џон Милтон, реченициите имале должина од дури 52 зборови.

- Анализирано по страници, страниците во кои имаме значајни отстапувања (outliers) во однос на просекот на искористени зборови по реченица, најчесто се страници во кои има само неколку реченици.



Должина на зборовите во книгата

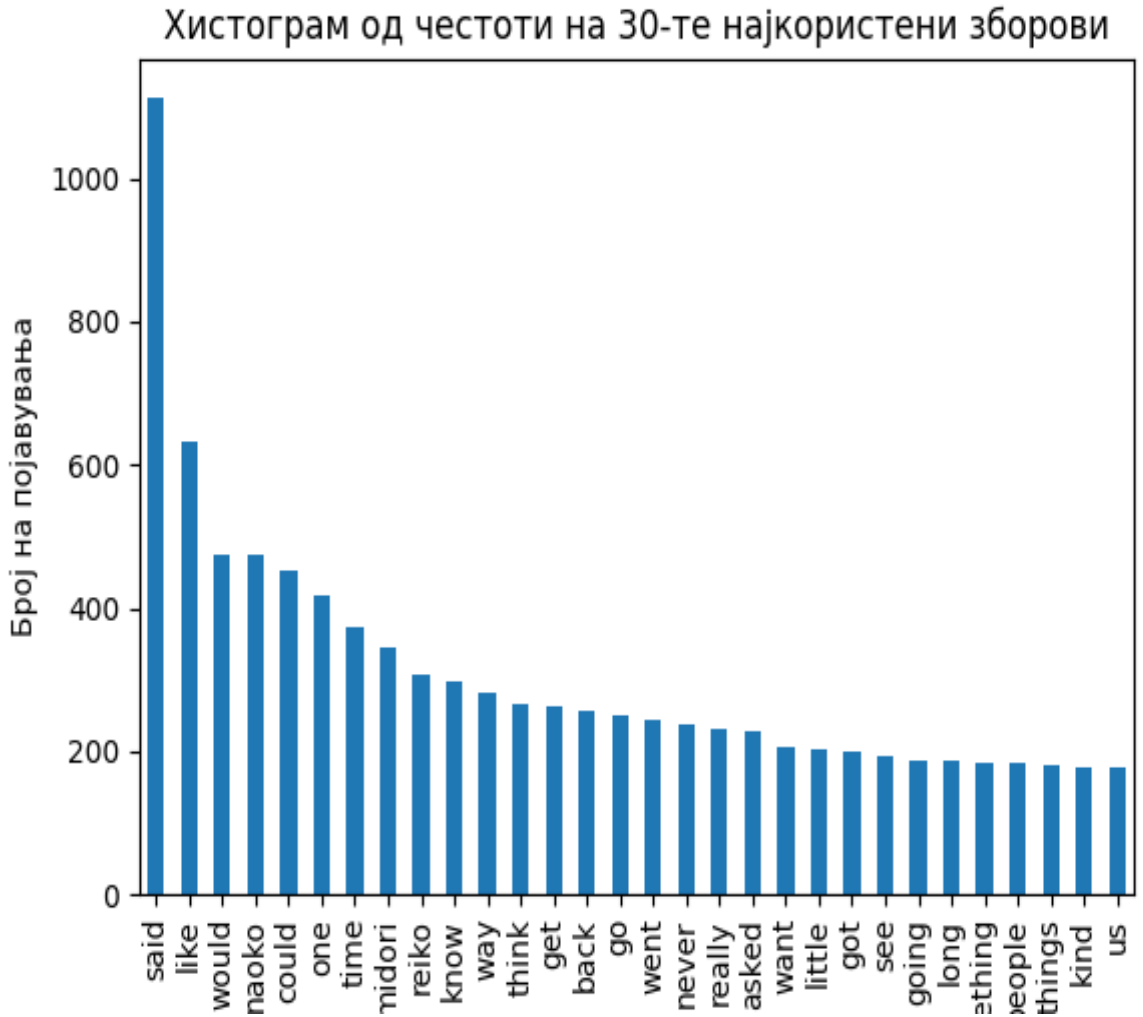
- Просечниот број на букви во зборовите од „Норвешка шума“ е околу 4, што е и вообичаен број за повеќето книги (околу 4 до 5 букви).



Честота на појавување на зборовите во книгата

Препроцесирање на податоците:

- сите букви во зборовите да бидат мали или големи букви.
- отфрлање на зборови за кои се смета дека не менуваат ништо во смислата на реченицата (пр. English stopwords колекцијата во Python ги отфрла зборовите како the, of, a, this, is, it, to…).
- морфолошка анализа на зборовите и кратење (сведување) на зборовите на нивните корени (пр. Reading и read имаат иста основа (read)).

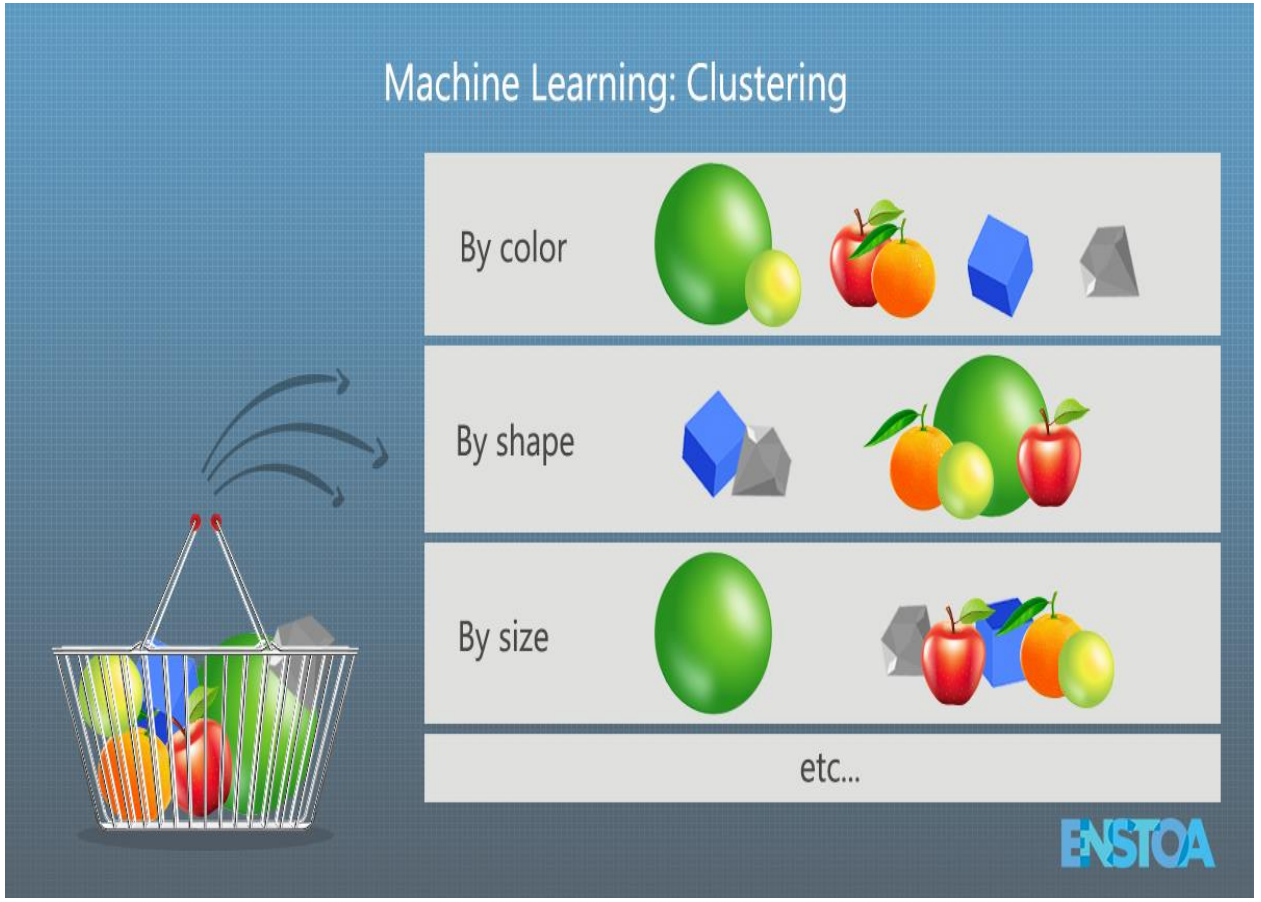


Машинско учење без набљудување (unsupervised learning)

- алгоритмите од машинското учење се применуваат врз множество податоци кое нема зависна (таргет) променлива.

- се користат за откривање „скриени патерни“ во податоците, откривање групи (кластери) кои се слични на некаков начин, распределбата на податоците во просторот итн.

- кластерирањето е метод на групирање на податоците во групи, така што податоците во една група имаат поголеми сличности (се поблиски, според дадена метрика) со податоците од својата група, отколку со оние во другите групи.



Кластерирање на страниците од „Норвешка шума“

- препроцесирање на податоците (сите букви претворени во мали букви и отфрлање на зборовите кои влегуваат во python класата на english stopwords).
- како предиктори (input променливи) може да се користат честотите на појавување на секој од зборовите во секоја од страниците.
- покрај честотите на појавување на зборовите, како предиктори може да се користи и tf-idf (term frequency – inverse document frequency) мерата на зборовите, која се користи во машинското учење.

$$tf - idf = tf(t,d) * idf(t)$$

tf(t,d) – релативниот број на појавувања на поимот (зборот) t во документот (на страницата) d во однос на вкупниот број на поими во документот d.

$$idf(t) = \log\left(\frac{n}{df(t)}\right)$$

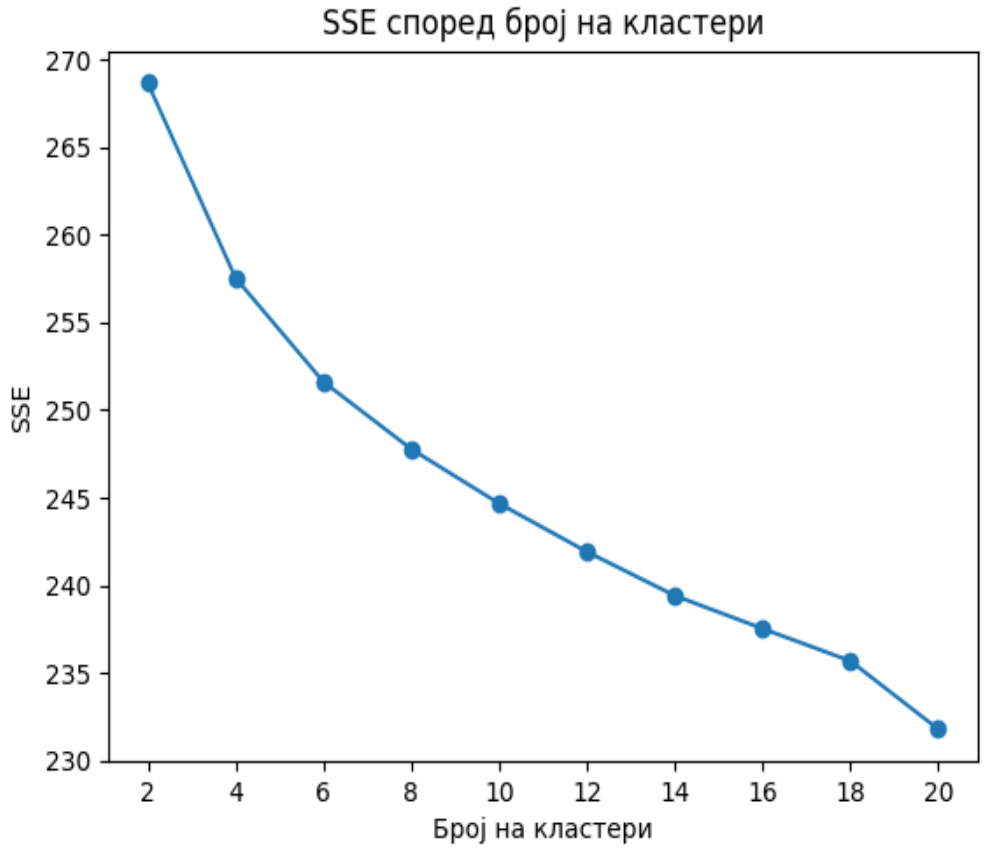
n – вкупен број на документи (на страници).
df(t) – број на документи (на страници) на кои се појавува поимот t.

Во модулите на Python, idf(t) се пресметува како:

$$idf(t) = \log\left(\frac{1+n}{1+df(t)}\right) + 1$$

Кластерирање на страниците од „Норвешка шума“ со користење на Kmeans алгоритамот

- во алгоритамот како влезни (input) податоци користиме 347 податоци (страници) на кои одговараат 231 колона со предиктори (tf-idf мерата на 231 одбрани зборови од книгата).
- одбравме да бројот на кластери k биде 6 (најчесто се користи „elbow“ методот).
- се користи евклидска норма како метрика за оддалеченост на секој податок од центроидот на кластерот.



Кластерирање на страниците од „Норвешка шума“ со користење на Kmeans алгоритамот

Десет најчесто користени зборови по кластери:

Кластер 0: 'nagasawa', 'like', 'one', 'know', 'think', 'girls', 'time', 'girl', 'go', 'people‘.

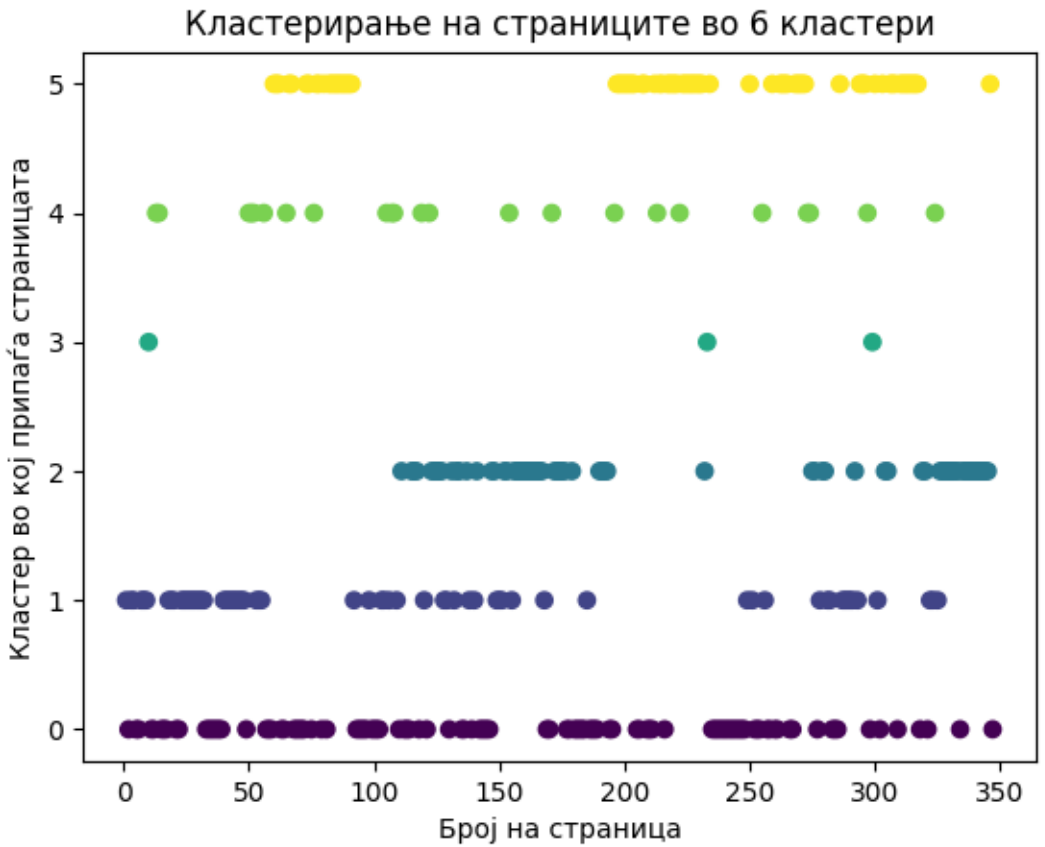
Кластер 1: 'naoko', 'kizuki', 'never', 'time', 'one', 'back', 'nothing', 'wer 'wanted', 'felt‘.

Кластер 2: 'reiko', 'naoko', 'like', 'one', 'time', 'us', 'asked', 'came', 'bac 'good‘.

Кластер 3: 'letter', 'dorm', 'lot', 'day', 'two', 'live', 'university', 'either', 'flat', 'another‘.

Кластер 4: 'room', 'window', 'small', 'back', 'stood', 'like', 'left', 'dark', 'naoko', 'night‘.

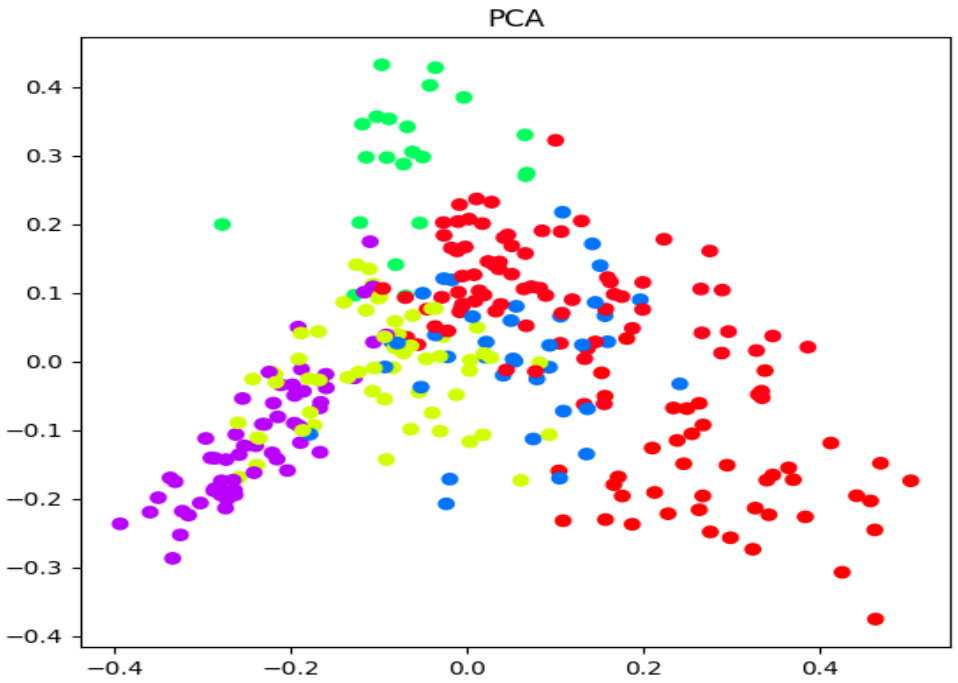
Кластер 5: 'midori', 'like', 'father', 'know', 'really', 'love', 'get', 'think', 'little', 'want‘.

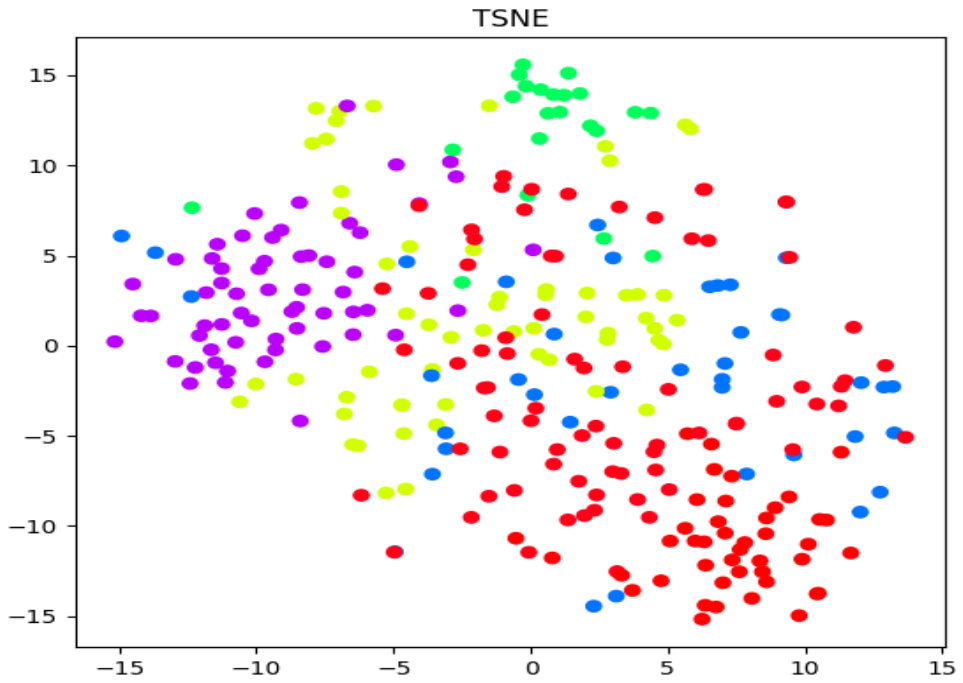


Кластерирање на страниците од „Норвешка шума“ со користење на Kmeans алгоритамот

PCA (Principal component analysis) – метод со кој се редуцира димензијата на вектор со повеќе компоненти (во нашиот случај, секој предиктор, кој е вектор со 231 компонента, се намалува на вектор со 2 компоненти), но и покрај намалувањето на димензијата, се задржуваат првичните својства на векторот.

TSNE (t-distributed stochastic neighbor embedding) – метод за редукција на димензијата кој на секоја точка и доделува перцентил од студентова распределба, кој работи добро и при нелинеарна зависност меѓу податоците.



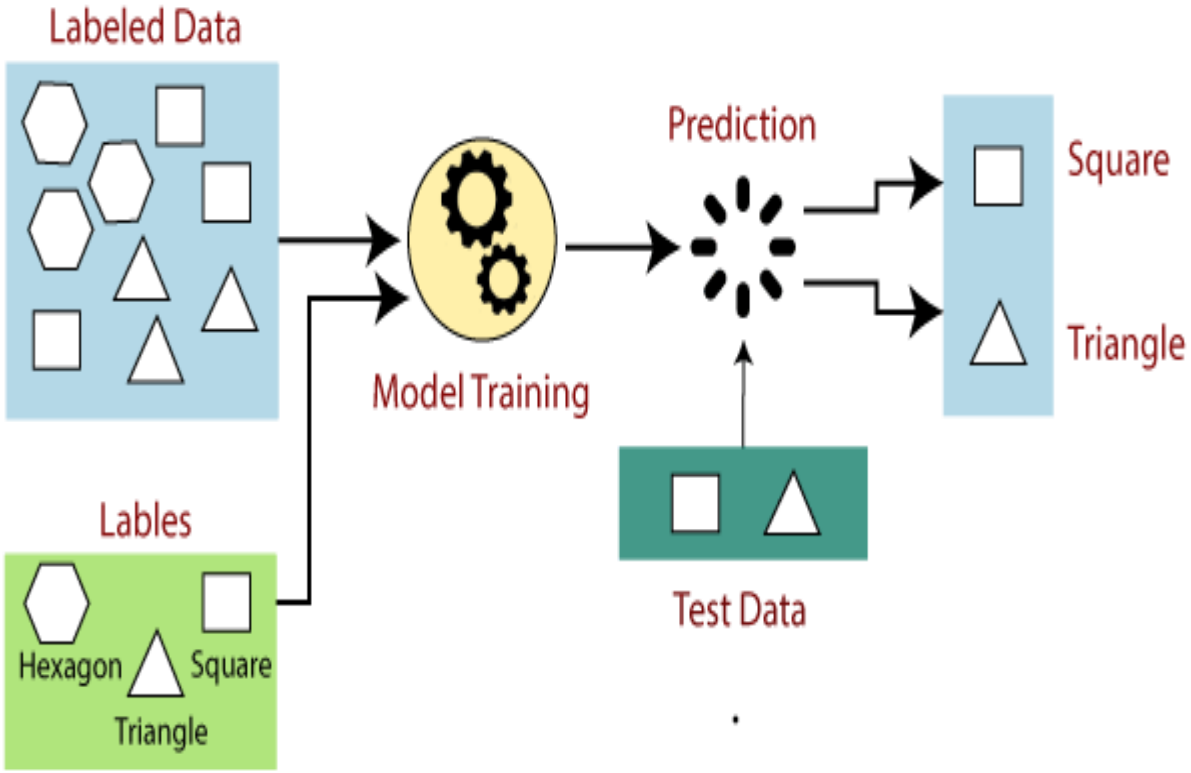


Машинско учење со набљудување (supervised learning)

- алгоритмите од машинското учење се применуваат врз множество податоци во кое покрај независните променливи се дадени и вредностите на зависната (таргет) променлива.

- се користат за класификација на податоците кога зависната променлива е категориска променлива или предвидување на вредноста на зависната нумеричка променлива.

- логистичката регресија е метод за класификација на податоците во случај кога зависната променлива е категориска, во кој класификацијата се врши со помош на предвидени веројатности.



Предвидување на оценката за „Норвешка шума“ според текстот од критичкиот осврт

- множество од 77 критички осврти (reviews) и оценките кои читателите ги дале за книгата, преземени од goodread.com.
- предвидување на оценката според критичкиот осврт, користејќи логистичка регресија, при што како независни променливи се tf-idf мерите на зборовите (земени се предвид честотите на единечните зборови и на паровите од два последователни зборови) од препроцесираниот текст од критичките осврти.
- во логистичката регресија употребуваме и L_2 регуларизација.
- точноста на моделот е околу 69% (толкав процент од критиките биле предвидени со точна оценка).