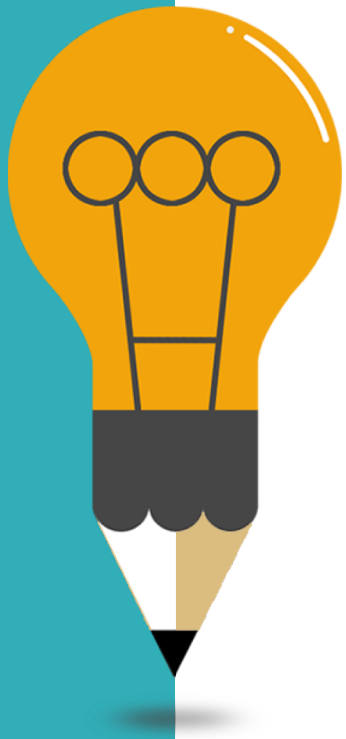
TEXT DOCUMENT CLUSTERING

Lubna Alhenaki

Outline



Problem Statement

Solution?

Clustering

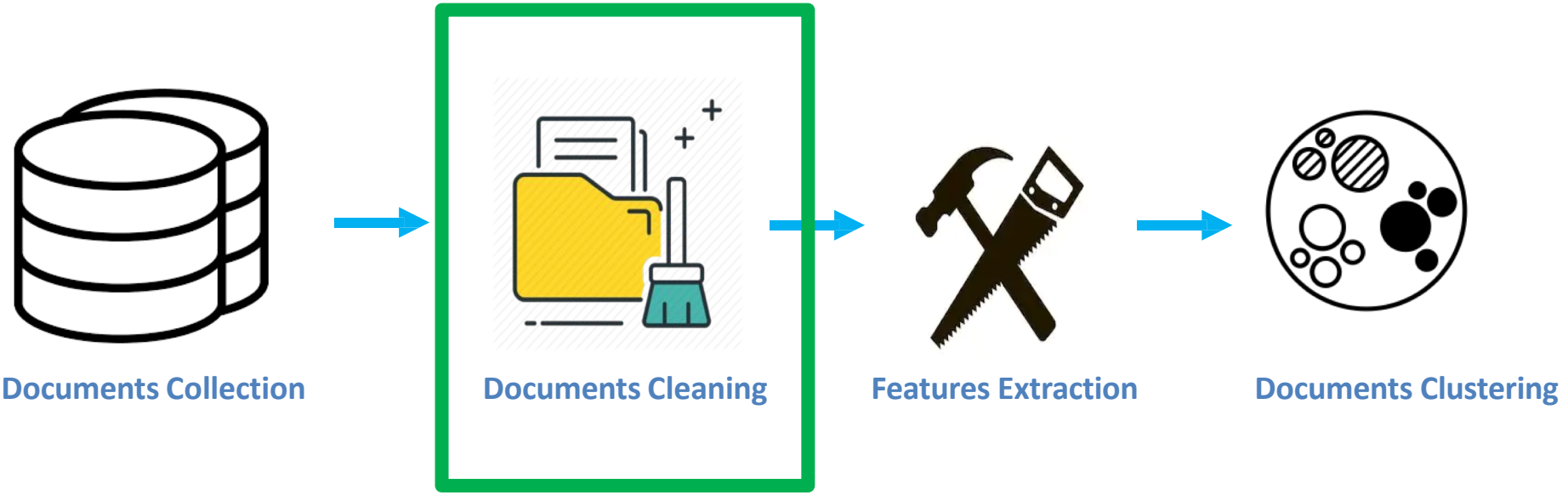
Text Document Clustering

- Text document clustering: Grouping of text documents into meaningful clusters in an unsupervised manner.
- Cluster Hypothesis : Relevant documents tend to be more similar to each other than to non-relevant ones.

Documents Clustering Methodology Design

Description of **Text Document Dataset**

Documents Clustering Methodology Design



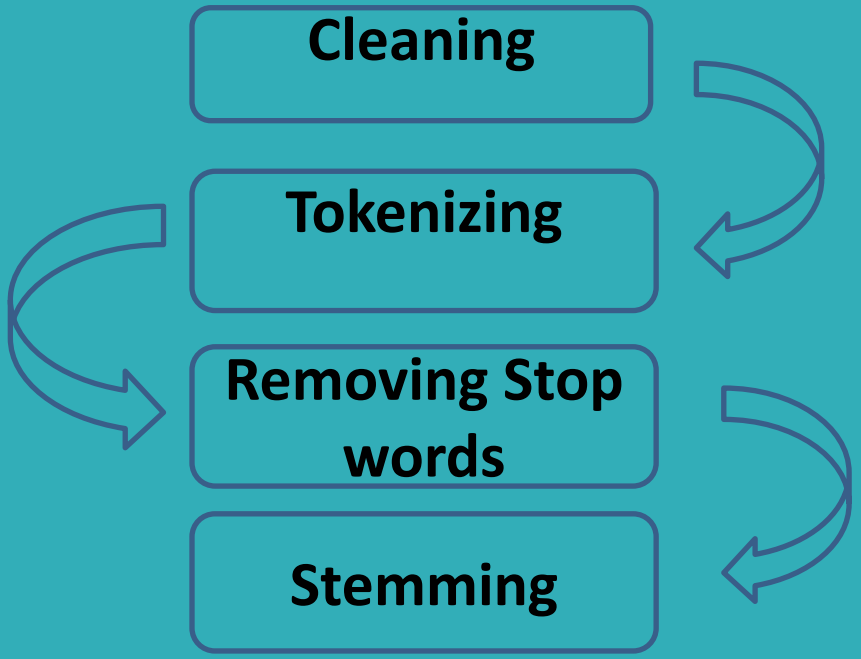
Pre-Processing Steps

Cleaning :

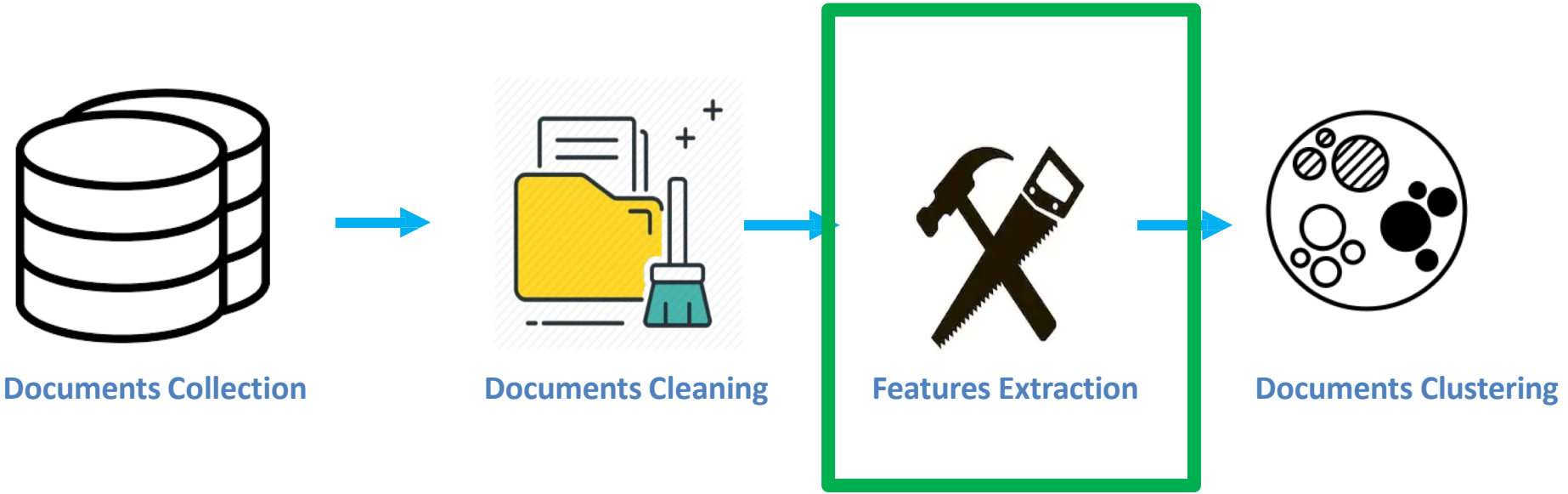
01 Make all text lower case

02 Remove punctuation and special character

03 Remove numerical values



Documents Clustering Methodology Design



Dimensiona Reduction / Topic Modeling

Baseline

- Vectorize using TFIDF (min_df= 500)
- Normalize (default by TFIDF)
- SVD (# of component=20)

Experiment 2

- Vectorize using TFIDF (min_df=10)
- Normalize (default by TFIDF)
- SVD (# of component=20)

Experiment 3

- NMF (# of component=20)

Experiment 4

- Vectorize using TFIDF (min_df=10, max_df=.5)
- Normalize(default by TFIDF)
- NMF (# of component=20)

Documents Clustering Methodology Design

Documents Collection

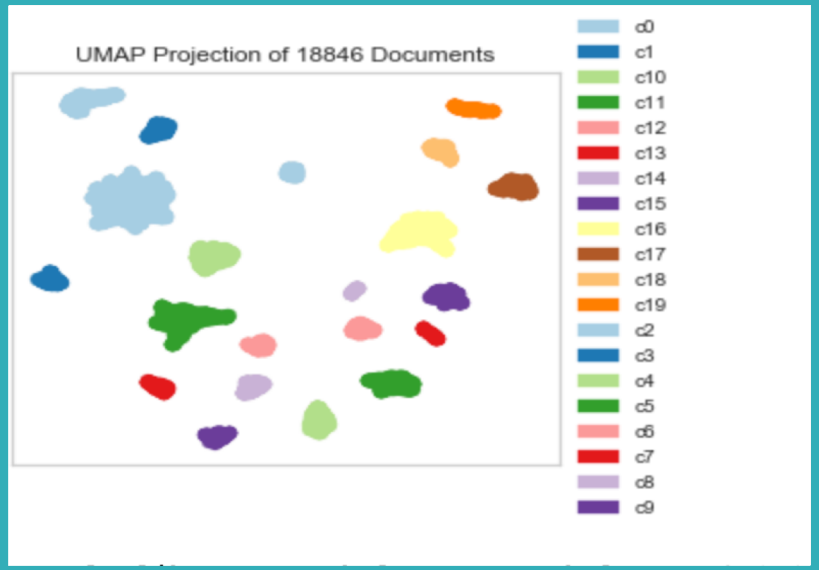
Documents Cleaning

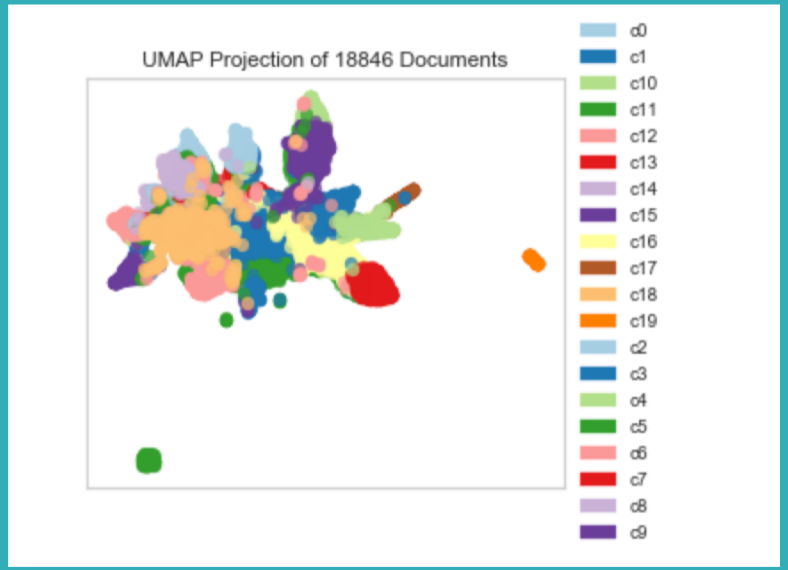
Features Extraction

Documents Clustering

Text Document Clustering

- **K-means Clustering Algorithm**
- **Number of clusters = 20**





Best Model Vs Baseline

```
In [277]:  1  pd.Series(cluster_model.labels_).value_counts()

Out[277]: 11    4781
          13    2024
          3     1896
          16    1606
          1      951
          18     924
          9      906
          5      745
          12     735
          7      592
          2      505
          15     474
          8      452
          4      451
          14     430
          6      426
          0      369
          10     343
          17     163
          19      73
```

Let’s inspect the results of clustering ☺

What is each cluster talking about?

0	(Topic16, 53.34)	(Topic0, 11.77)	(Topic18, 5.1)
1	(Topic17, 57.78)	(Topic9, 7.07)	(Topic0, 4.21)
2	(Topic18, 14.21)	(Topic0, 11.13)	(Topic16, 9.21)
3	(Topic12, 42.0)	(Topic19, 13.34)	(Topic15, 6.89)
4	(Topic0, 28.9)	(Topic4, 16.87)	(Topic18, 10.15)
5	(Topic18, 22.43)	(Topic1, 13.27)	(Topic19, 12.25)
6	(Topic3, 48.85)	(Topic15, 6.83)	(Topic1, 6.64)
7	(Topic6, 60.19)	(Topic16, 6.91)	(Topic0, 5.81)
8	(Topic4, 56.79)	(Topic0, 13.4)	(Topic14, 4.37)
9	(Topic10, 44.9)	(Topic1, 7.84)	(Topic14, 7.73)
10	(Topic2, 46.44)	(Topic0, 11.64)	(Topic14, 5.2)
11	(Topic9, 32.31)	(Topic15, 17.45)	(Topic5, 8.57)
12	(Topic7, 48.19)	(Topic18, 7.38)	(Topic10, 7.31)
13	(Topic11, 74.99)	(Topic16, 4.72)	(Topic0, 3.98)
14	(Topic13, 90.97)	(Topic0, 2.18)	(Topic1, 1.16)
15	(Topic14, 42.69)	(Topic5, 8.09)	(Topic15, 8.03)
16	(Topic0, 20.72)	(Topic17, 14.98)	(Topic18, 11.82)
17	(Topic5, 47.48)	(Topic14, 11.82)	(Topic18, 7.49)
18	(Topic2, 70.8)	(Topic0, 5.61)	(Topic9, 3.58)
19	(Topic8, 60.72)	(Topic0, 6.82)	(Topic11, 5.97)

Example

11	(Topic9, 32.31)	(Topic15, 17.45)	(Topic5, 8.57)
12	(Topic7, 48.19)	(Topic18, 7.38)	(Topic10, 7.3
13	(Topic11, 74.99)	(Topic16, 4.72)	(Topic0, 3.98)
14	(Topic13, 90.97)	(Topic0, 2.18)	(Topic1, 1.16)

```
 ** Topic 7
file, format, program, ftp, convert, gif, directori, zip, imag, disk,

 ** Topic 8
israel, arab, jew, isra, jewish, palestinian, state, kill, peac, war,

 ** Topic 9
sale, price, offer, new, sell, ship, condit, includ, ask, best,
```

```
In [304]:  1  documents[49]
```

```
Out[304]: 'i am looking for the exact address of the symantec coporatoin  which \ndistributes norton desktop and other windows
software \n\nthe information i am looking for is \n\nmail address\nphone number\nfax number\ne mail address\n\nthanks
in advance \n'
```

Conclusion and Future Work

Text document clustering groups similar documents to form a coherent cluster, while documents that are different have separated apart into different clusters.

In this project, k-means algorithm applied into 20-newsgreoup dataset and get approximately good clustering results.

For Future work, features selection and other popular text dataset su ch as Reuters dataset could be used. Furthermore, other algorithm such as Genetic algorithm

Thank You

Any Questions?

Appendix:

```
 ** Topic 0
think, peopl, say, thing, go, whi, make, see, want, good,

 ** Topic 1
problem, work, run, tri, system, machin, fix, time, error, fine,

 ** Topic 2
game, team, play, year, player, win, hi, hockey, season, score,

 ** Topic 3
drive, scsi, disk, hard, ide, floppi, control, boot, meg, hd,

 ** Topic 4
god, christian, jesu, hi, believ, bibl, sin, christ, faith, church,

 ** Topic 5
pleas, mail, post, list, address, email, send, repli, thank, edu,

 ** Topic 6
key, chip, encrypt, clipper, escrow, secur, phone, bit, algorithm, govern,

 ** Topic 7
file, format, program, ftp, convert, gif, directori, zip, imag, disk,

 ** Topic 8
israel, arab, jew, isra, jewish, palestinian, state, kill, peac, war,

 ** Topic 9
sale, price, offer, new, sell, ship, condit, includ, ask, best,
```

Appendix:

```
 ** Topic 10
window, run, program, ms, applic, os, font, manag, version, app,

 ** Topic 11
armenian, turkish, muslim, genocid, armenia, turk, turkey, peopl, russian,

 ** Topic 12
card, driver, video, bu, diamond, ati, vlb, isa, mode, slot,

 ** Topic 13
edu, geb, dsl, cadr, chastiti, pitt, bank, intellect, gordon, surrend,

 ** Topic 14
anyon, thank, doe, advanc, appreci, hi, help, info, could, inform,

 ** Topic 15
modem, port, mac, pc, serial, printer, connect, softwar, pin, board,

 ** Topic 16
gun, law, govern, fbi, right, fire, batf, state, weapon, koresh,

 ** Topic 17
car, bike, engin, ride, mile, buy, look, speed, front, oil,

 ** Topic 18
space, system, program, inform, orbit, data, nasa, avail, comput, also,

 ** Topic 19
monitor, color, imag, display, screen, bit, vga, video, appl, graphic, CPU
```