Name: Farhan Ali
Course: COMP 479 – Information Retrieval
Assignment: Crawling, Indexing, TF–IDF, and Clustering

1. Purpose of This Demo

This Demo file briefly walks through:

- how the clustering was run,
- examples of good clusters
- examples of bad clusters
- a short comparison between global DF and local DF.

This file is meant to show that the system works and that I understand the behaviour of the clusters.

2. How the Clustering Was Run

After crawling 300 PDFs from Spectrum and creating My-collection (54 documents), I ran clustering using:

```
python experiments_clustering.py
```

Inside the file, two runs happen:

- `main(df_mode="global")` → required
- `main(df_mode="local")` → bonus

For each DF mode, the script creates a TF–IDF matrix and applies K-Means with:

- **k = 2**
- **k = 10**
- **k = 20**

Each cluster prints:

- cluster size
- top terms
- 3 example documents

The examples below come directly from those outputs.

3. Good Cluster Examples

3.1 k = 2 – Strong Natural Split

With **global DF**, k = 2 gives a very clear high-level split.

Cluster A Sustainability / Social Topics (96 docs)
Top terms:
`sustain, person, commun, offer, question, social, learn, engag, peopl, servic, opportun, beyond, benefit, way, issu, futur, individu, need, focu, resourc`

Why good:

- Terms describe sustainability, community engagement, education, and social/leadership work.
- Documents include leadership PDFs, sustainability reports, and social development papers such as `Azzizi_MCompSC_F2025.pdf`, `Table I Grief Leadership.pdf`, and the upcycling dataset.

Cluster B Technical / Experimental / Scientific (68 docs)
Top terms:
`sustain, ga, ratio, mm, energi, paramet, temperatur, diamet, condit, carri, variat, laboratori, locat, industri, figur, size, fig, averag`

Why good:

- Highly technical vocabulary: measurements, parameters, temperature, figures, lab context.
- Contains biology, ecology, and experimental papers like `Zerges_JCB1998.pdf`, `Bachewich_JCS1998.pdf`, and `dayanandan-2016-Ecology_and_Evolution.pdf`.
- Clear separation between **community/sustainability language** and **lab/measurement language**.

This is a strong example of K-Means capturing two major “styles” or disciplines in the dataset.

3.2 k = 10 – Clear Academic Subfields

With **k = 10 (global DF)**, several very coherent academic subfields appear.

Combustion / Detonation Cluster
Top terms:
`deton, flame, kpa, stoichiometr, hoi, ignit, combust, cj, unstabl, instabl, shock, transduc`

Example docs:
`hydrogen_gao.pdf`, `POF.pdf`, `paper_porous_hd.pdf`

Why good:
Extremely coherent combustion/detonation vocabulary → a clear engineering subfield centred on ignition, shock waves, and explosives.

Neuroscience Cluster
Top terms:
`pharmacolog, prefront, neurobiolog, nucleu, dopamin, receptor, rat, neurosci, hippocampu, accumben, cortex`

Example docs:
`Amir-f1000R-2016.pdf`, `fnbeh-10-00238.pdf`, `chapman-pone-2015.pdf`

Why good:
All terms relate to brain regions, neurobiology, behaviour, and pharmacology. The papers are clearly grouped by neuropsychopharmacology content.

CFD / Wind / Pollution Cluster
Top terms:
`cfd, stathopoulo, turbul, veloc, ashra, tunnel, rooftop, wind, aerodynam, rough, roof, intak, emiss`

Example docs:
`Ten Questions Concerning Modeling of Near-Field Pollutant Dispersion...pdf`, `SE-D-16-00202R2.pdf`, `HE-D-16-00340R1.pdf`

Why good:
Very specific fluid dynamics and wind engineering vocabulary. Only the CFD / rooftop dispersion papers appear here.

Overall, **k = 10** gives some of the strongest and most interpretable clusters.

4. Bad / Weak Cluster Examples

4.1 Singleton Clusters (k = 10 / k = 20)

Example singleton cluster (k = 10, global DF):

- **Size:** 1
- **Top terms:** unusual author/method names like
  `cessac, latora, aihara, amari, microst, cybern, ...`
- **Example doc:** `InterneuronDynamics-IJBC2.pdf`

Why bad (for interpretation):

- Only one document → no group structure to compare or label.
- This happens because K-Means is forced to create 10 (or 20) clusters even when the data naturally has fewer tight groups.
- Shows **over-fragmentation**, a classic limitation of K-Means when k is too large.

4.2 Vague Social Cluster (k = 20)

Example social cluster (k = 20, global DF):

- **Top terms:**
  `voic, sustain, stori, inquiri, talk, ethic, bring, live, peopl, came, articul, felt, listen, educ, engag`

Why weak:

- Terms are generic and span multiple social-science topics (voice, story, ethics, education, feelings).
- The cluster is clearly “social / narrative”, but it is harder to give it a precise label compared to very sharp clusters like combustion or neuroscience.
- This demonstrates cases where clustering becomes fuzzy and harder to interpret as k increases.

5. Global DF vs Local DF

Global DF

- Computes document frequency across **all 998 crawled documents**.
- Emphasizes rare scientific terms like:

    - combustion vocabulary (`deton, kpa, stoichiometr, ignit`)
    - neuroscience vocabulary (`dopamin, hippocampu, cortex, accumben`)
- Leads to **very sharp scientific clusters** (combustion, neuroscience, CFD, membranes, ecology).

Local DF

- Computes DF only across the **164 documents in My-collection**.
- Emphasizes terms that are rare within this subset, such as:
    - economic and service terms (`econom, cost, servic, busi`)
    - narrative/social terms (`stori, friend, youth, experienc`)
    - librarianship / French-language terms in smaller clusters.
- Makes sustainability/social clusters more fine-grained and slightly reorganized, while scientific clusters (combustion, neuroscience, CFD) stay almost the same.

Conclusion:

- Scientific clusters remain very similar in both modes because their vocabulary is highly specialized.
- Social/sustainability clusters shift more between global and local DF because their vocabulary is broader and more varied.

6. Summary

This demo shows that:

- The IR pipeline works end-to-end:
  crawl → extract → index → query → cluster.
- **k = 2** reveals broad “social/sustainability vs technical/experimental” themes.
- **k = 10** produces the best, most meaningful academic subfields (combustion, neuroscience, CFD, ecology, etc.).
- **k = 20** exposes very fine sub-topics but also creates noisy or tiny clusters (including singletons).
- **Global vs Local DF** changes which terms dominate, especially inside the sustainability/social part of My-collection, while technical clusters remain very stable.

Overall, the system produces clear, interpretable clusters for small/medium k, and reveals natural limitations when k is too large for the dataset