# Israeli Media Monitor — Latest Questions (Summary)

*Generated: 2025-08-28 16:41*

## Normalization

* What do you mean by “normalized versions” of title + summary?
* In normalization, should I remove symbols like •, :, ", the pipe and suffix ("| N12"), the prefix "צפו:", hyphens, etc.?
* Given my current norm\_min function, should I add these extra rules—and in what order?
* Explain what each line in the normalizer does.
* Is there an easier way—like keeping only Hebrew letters? Pros/cons?
* My text\_for\_cluster isn’t normalized—is that OK or should I normalize or normalize inside the vectorizer?

## Vectorization & Stopwords

* Is my TfidfVectorizer word-based by default?
* How can I load Hebrew stopwords from a TXT file (analysis/utils/hebrew\_stopswords\_list\_extended.txt) and pass them into TfidfVectorizer?
* Should I also use a character-level TF-IDF? What’s the difference between word vs char TF-IDF and when to use each?

## Clustering Method

* How do I start clustering from the beginning (simple, step-by-step)?
* Can you give me a minimal one-file script for clustering (group\_similar.py)?
* In AgglomerativeClustering, should the metric be 'precomputed' when I pass a distance matrix?
* What should the clustering represent for news articles (i.e., 'same real-world event')?
* Provide example titles + summaries that should land in the same cluster.

## Saving & Inspecting Results

* Can I save outputs to CSV / JSON — which is better for inspection?
* How should I examine the results to see if clustering worked (clusters\_summary, articles.csv, JSONL, quick REPL checks)?
* How to compute quick diagnostics (cross-outlet mix, top terms per big cluster) and where to run them?

## Quality & Size Questions

* Is 34 stories enough, or should I have more? A bit more or much more?
* How many singletons are too many? What’s a healthy range?
* What’s a good target for cross-outlet clusters (clusters with ≥2 sources; % articles in those clusters)?

## CLI & Parameters

* What is argparse.ArgumentParser exactly?
* Is my argparse/main block OK (processed-dir, threshold, min-df, max-df, ngrams, max-articles, out-dir, save)?
* I’ve switched to title\_norm\_min and summary\_norm\_min to build text\_for\_cluster — is that better than using raw title + summary?