01 🖍️ Armenian Alphabet Entropy

The Armenian alphabet has **39 letters** (Mashtots' original 36 + ՈՒ, Օ, Ֆ).

a) Compute the **uniform entropy**: how many bits per letter if every letter were equally likely?

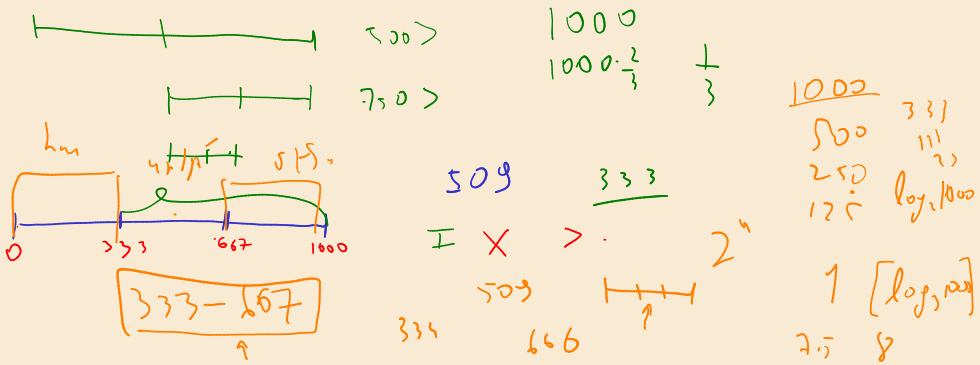
b) English's uniform bound is $\log_2 26 \approx 4.70$ bits/letter, yet real English text has empirical entropy around 1.3 bits/letter (Shannon, 1951). Explain *why* the empirical entropy is so much smaller than the uniform bound. Name the **two** distinct effects.

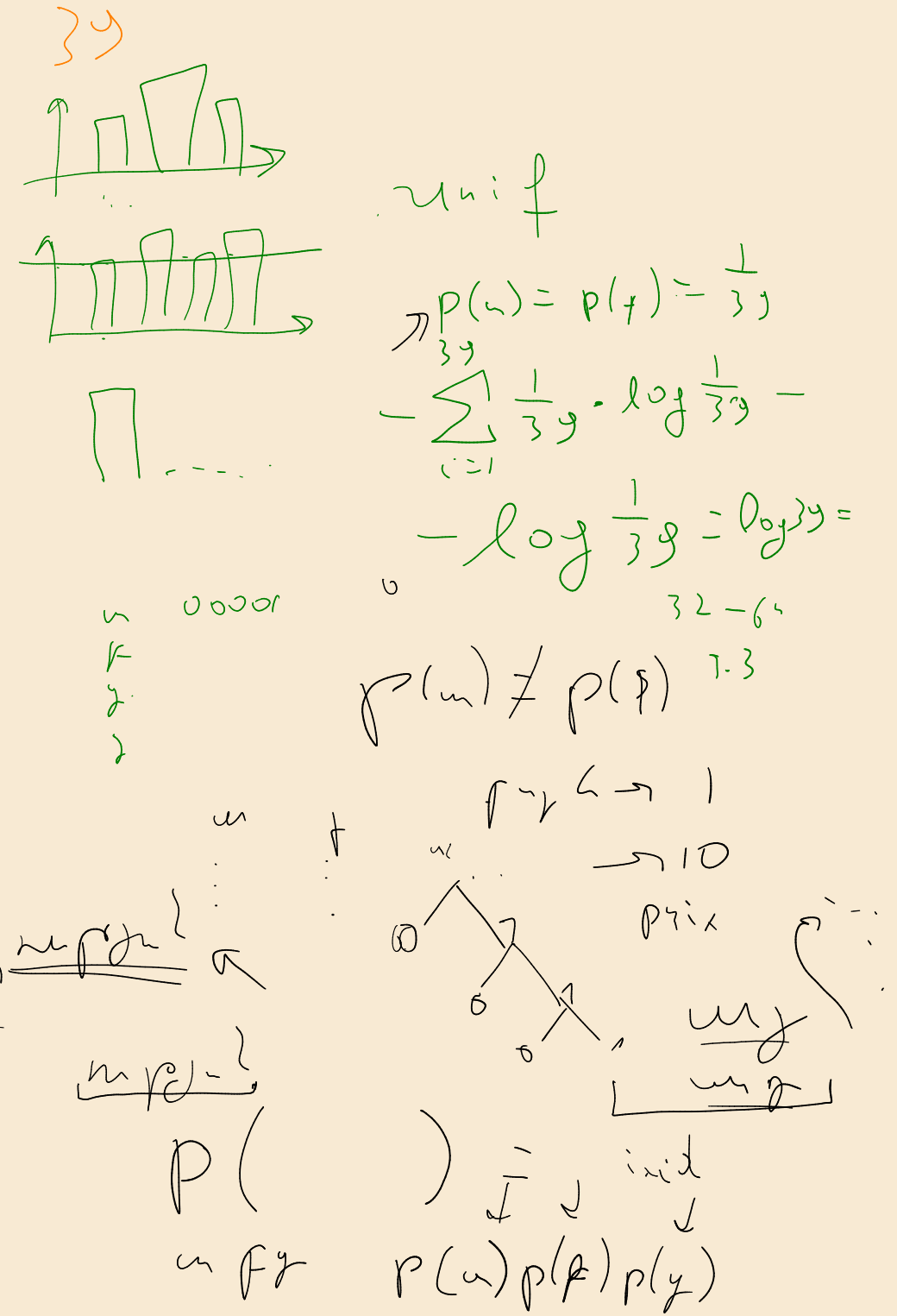
c) (Python) Estimate the *real* per-letter entropy of Armenian. Use the provided file `assets/panir_hy.txt` (the «Պանիր» Wikipedia article), or download fresh text yourself with the snippet below. Keep only Armenian letters, lowercase them, and compute the empirical letter-frequency distribution and its entropy.

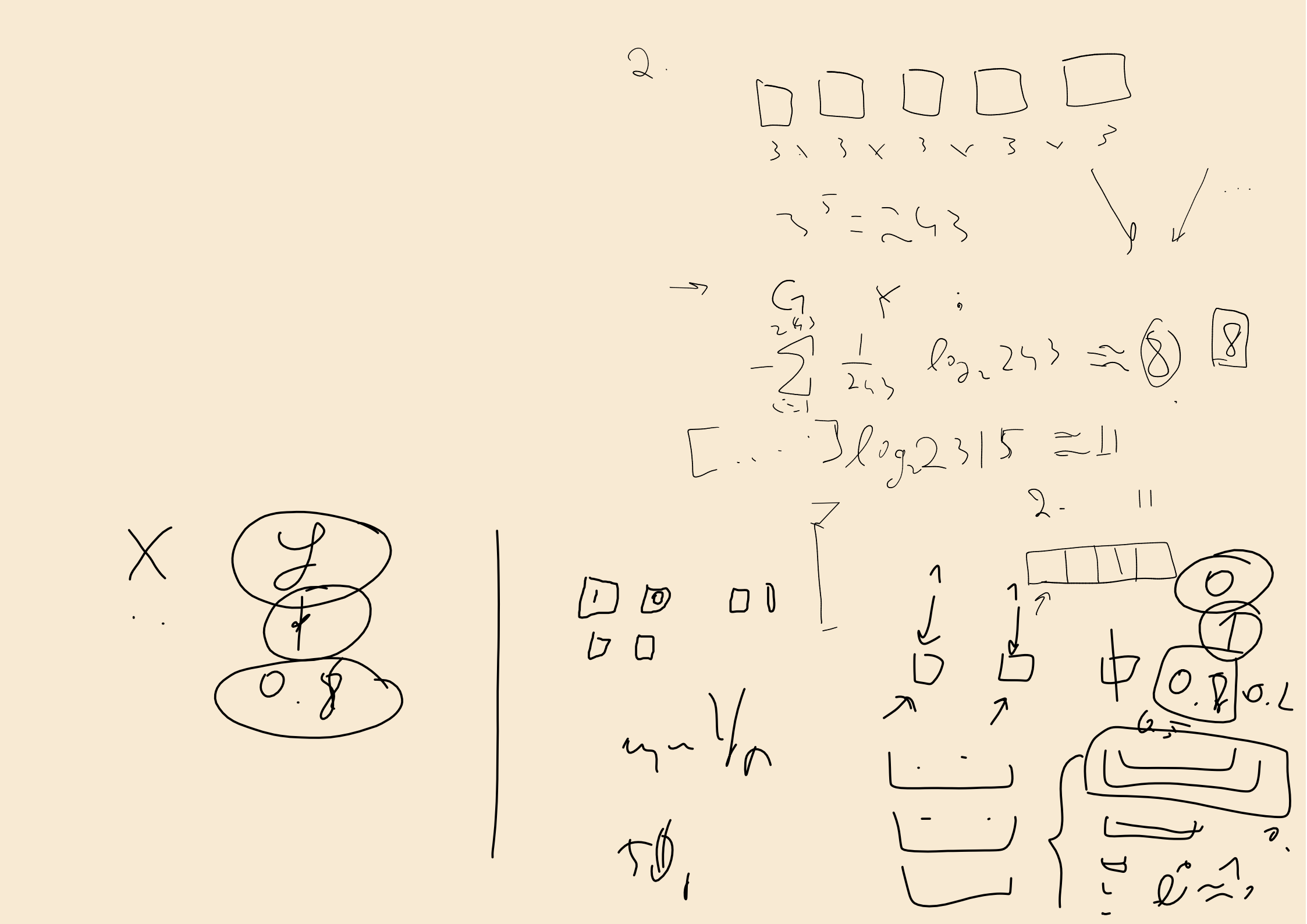
```
# Option A — download the provided text file:
import urllib.request
url = ("https://raw.githubusercontent.com/HaykTarkhanyan/"
       "python_math_ml_course/main/math/assets/panir_hy.txt")
text = urllib.request.urlopen(url).read().decode("utf-8")

# Option B — download any Armenian article's plain text yourself:
import urllib.request, urllib.parse, json
title = "Պանիր"  # try any article title
url = "https://hy.wikipedia.org/w/api.php?" + urllib.parse.urlencode({
    "action": "query", "prop": "extracts", "explaintext": 1,
    "titles": title, "format": "json", "redirects": 1,
})
req = urllib.request.Request(url, headers={"User-Agent": "course/1.0"})
page = next(iter(json.load(urllib.request.urlopen(req))["query"]["pages"].values()))
text = page["extract"]
```

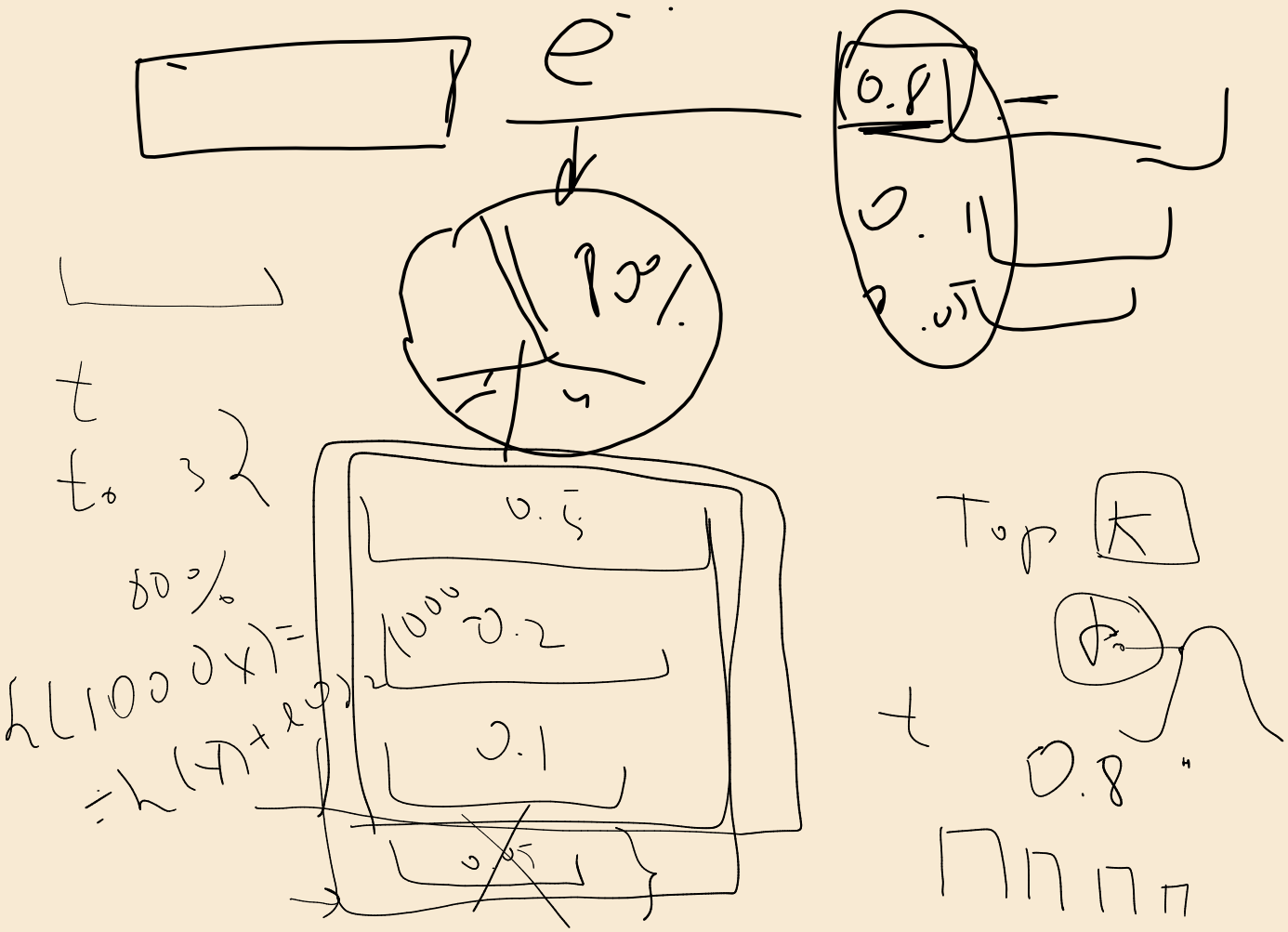
d) Compare your empirical entropy to the uniform bound from (a). By roughly how many bits per letter does the frequency skew save you?

e) A 100-page Armenian e-book has roughly $200{,}000$ characters. Estimate the entropy **lower bound** on its size (KB). Then explain the catch: you'll get ≈ 4.5 bits/letter, but a real code must assign a **whole** number of bits to each letter — so how do you actually get close to the entropy?








e⁻
0.8
0.1
0.05
0.5
0.2
0.1
0.05
80%
Top