1. domača naloga: Nedoločenost naravnega jezika

Claude E. Shannon je leta 1948 v prelomnem članku *A Mathematical Theory of Communication* [1] postavil temelje moderne informacijske teorije. V njem je vpeljal nedoločenost (entropijo) kot mero informacije. V članku *Prediction and Entropy of printed English* [2], iz leta 1951, je omenjeno mero uporabil pri analizi angleških besedil. Hotel je ugotoviti, koliko informacije nosi v povprečju posamezna črka. Problema se je lotil analitično in eksperimentalno. Rezultati so pokazali, da črka angleške abecede nosi v povprečju približno 1,5 bita informacije [3].

Podoben poskus želimo ponoviti tudi mi. Ugotoviti želimo približek povprečne informacije, ki jo nosijo črke angleške abecede. Izhajamo iz nedoločenosti naključne spremenljivke X, ki predstavlja črko angleške abecede. Ob predpostavki, da so zaporedne črke neodvisne in poznamo njihove verjetnosti, velja:

$$H(X) = -\sum_{i}^{I} p(x_i) \log_2 p(x_i) \quad . \tag{1}$$

Če z naključnima spremenljivkama X_p in X_n označimo dve zaporedni črki v besedilu in zgornja predpostavka drži, lahko za nedoločenost para črk zapišemo:

$$H(X_p, X_n) = H(X_p) + H(X_n) \quad . \tag{2}$$

Ker so črke v besedilu med seboj odvisne velja $H(X_p, X_n) < H(X_p) + H(X_n)$. Povprečno informacijo $H(X_n|X_p) < H(X_n)$, ki jo dobimo, ko izvemo črko X_n, lahko izračunamo kot

$$H(X_n|X_p) = H(X_p, X_n) - H(X_p) \quad . \tag{3}$$

Ta količina nam predstavlja povprečno informacijo na črko, ki jo dobimo, če predhodno črko že poznamo. Pri obravnavi nam X_p lahko predstavlja poljuben niz črk (pare, trojice, ...)

$$\mathbf{X_p} = (X_1, X_2, ..., X_{n-1}) \quad . \tag{4}$$

Ker X_n predstavlja naslednjo črko v istem besedilu, lahko zapis poenostavimo tako:

$$\mathbf{X_{pn}} = (X_1, X_2, ..., X_p, X_n) \quad , \quad \text{sledi} \tag{5}$$

$$H(X_n|\mathbf{X_p}) = H(\mathbf{X_{pn}}) - H(\mathbf{X_p}) \tag{6}$$

$$= -\sum_{i}^{I} p(\mathbf{x_{pn_i}}) \log_2 p(\mathbf{x_{pn_i}}) + \sum_{j}^{J} p(\mathbf{x_{p_j}}) \log_2 p(\mathbf{x_{p_j}}) \quad . \tag{7}$$

- I predstavlja število vseh možnih nizov dolžine n,
- J predstavlja število vseh možnih nizov dolžine $n-1$,
- $p(\mathbf{x_{pn_i}})$ je verjetnost pojavitve niza $\mathbf{x_{pn_i}}$ dolžine n,
- $p(\mathbf{x_{p_i}})$ je verjetnost pojavitve niza $\mathbf{x_{p_i}}$ dolžine $n-1$.

V limiti, ko se n približuje neskončnosti, se $H(X_n|\mathbf{X_p})$ približuje pravi vrednosti (povprečni informaciji na znak).

Naloga

V datoteki `naloga1.py` v jeziku Python napišite funkcijo z imenom `naloga1`, ki izračuna približek povprečne informacije na znak $H(X_n|\mathbf{X_p})$ za dano število poznanih predhodnih črk.

- Vhodna argumenta funkcije sta niz z besedilom in število poznanih predhodnih črk p (celoštevilska vrednost na intervalu [0,3]).
- Funkcija naj vrne nedoločenost H v bitih.
- Besedilo lahko vsebuje vse črke angleške abecede (male in velike), števila in poljubna ločila ter presledke, tabulatorje ipd.

Prototip funkcije

```
def naloga1(besedilo: str, p: int) -> float:
    """ Izracun povprecne nedolocenosti na znak

    Parameters
    ----------
    besedilo : str
        Vhodni niz
    p : int
        Stevilo poznanih predhodnih znakov: 0, 1, 2 ali 3.
        p = 0: H(X1)
            racunamo povprecno informacijo na znak abecede
            brez poznanih predhodnih znakov
        p = 1: H(X2|X1)
            racunamo povprecno informacijo na znak abecede
            pri enem poznanem predhodnemu znaku.
        p = 2: H(X3|X1,X2)
            racunamo povprecno informacijo na znak abecede
            pri dveh poznanih predhodnih znakih.
        p = 3: H(X4|X1,X2,X3)
            racunamo povprecno informacijo na znak abecede
            pri treh poznanih predhodnih znakih.

    Returns
    -------
    H : float
        Povprecna informacija na znak abecede z upostevanjem
        stevila poznanih predhodnih znakov 'p'. V bitih.
    """

    H = float("nan")
    return H
```

Testni primeri

Na učilnici se nahajajo trije testni primeri besedil, za katere imate podane tudi povprečno informacijo na znak `H` (v bitih) za predpisano število poznanih predhodnih znakov `p`. Podatki so podani v obliki datotek `.json`. Priloženo imate tudi funkcijo `test_naloga1`, ki jo lahko uporabite za preverjanje pravilnosti rezultatov, ki jih vrača vaša funkcija. Pri testiranju svoje rešitve upoštevajte naslednje omejitve:

- rezultat je pravilen, če se od danega razlikuje za manj kot 10^{-3};
- izvajanje vašega programa je časovno omejeno na 30 sekund;
- uporabljati smete samo tiste pakete, ki so del standardne knjižnice Python 3.12 (https://docs.python.org/3.12/library/) in paketa `numpy` ter `scipy`. Na našem sistemu za preverjanje domačih nalog drugi paketi niso nameščeni.

Namigi

Predprocesiranje

Iz vhodnih podatkov odstranite vse znake, ki niso črke angleške abecede. Črke nato spremenite v velike (a $\rightarrow$ A), da jih s tem poenotite. Primer:

`'Danes bom naredil 1.  domaco nalogo!'` $\rightarrow$ `'DANESBOMNAREDILDOMACONALOGO'`

Računanje povprečne informacije na znak

Najprej je potrebno izračunati verjetnosti posameznih nizov črk glede na to, kateri približek računamo oziroma koliko predhodnih znakov poznamo. Ko imamo izračunane verjetnosti, uporabimo formulo za entropijo.

Uporabne funkcije in razredi

- `count`
- `filter`
- `set`
- `sum`
- `zip`
- `math.log2`
- `collections.Counter`
- `str.isalpha`, `str.join`, `str.upper`

Literatura

[1] C. E. Shannon: A mathematical theory of communication. Bell system technical journal, zv. 27, 1948.

[2] C. E. Shannon: Prediction and entropy of printed English. Bell Systems Technical Journal, zv. 30, str. 50–64, 1951.

[3] D.G. Luenberger: Information Science, Princeton University, str. 43-44, 2006.