Використання прихованих марковських моделей для декодування тексту

(1) **Структура українського алфавіту**

Підготуйте великий текст українською мовою з буквами одного регістру та видаленими знаками пунктуації. Вважаємо, що текст складається з $M = 34$ різних знаків - 33-ьох букв українського алфавіту і знаку відступу між словами (якщо включити апостроф, то буде $M = 35$ знаків).

Щоб виділити дві групи знаків в алфавіті, побудуємо ПММ $\lambda = (\mu, A, B)$ на основі послідовності $\{(X_n, Y_n)\}_1^T$, де:

- $T = 50000$ - кількість спостережень (об'єм тексту);
- $\{(X_n)\}_{n\geq 0}$ ланцюг Маркова з двома станами $N = 2$ з невідомою матрицею перехідних ймовірностей $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ та початковим розподілом $\mu = (\mu_1, \mu_2)$;
- матриця B містить умовні ймовірності $B_{ij} = P(Y_n = j | X_n = i)$, $i = \overline{1, N}, j = \overline{1, M}$.

Розв'яжіть задачу навчаня для ПММ і знайдіть найбільш ймовірні матриці A, B, що відповідають спостереженому тексту. В якості початкового наближення покладіть $\mu_i \sim 1/N$, $A_{ij} \sim 1/N$, $B_{ij} \sim 1/M$, не забуваючи, що матриці мають бути стохастичними.

Отримавши в результаті навчання матрицю B, віднесіть кожен знак $j = 1, \ldots, M$ до одного з класів 1 або 2, порівнявши ймовірності B_{1j} B_{2j}.

Які висновки можна зробити про структуру українського алфавіту на основі отриманих матриць A, B?

Спробуйте розділити знаки алфавіту на $N = 3$, $N = 4$ класи.

(2) **Декодування зашифрованого тексту**

Підготуйте текст з 50000 знаків українською мовою (видаліть знаки пунктуації, пропуски між словами, всі букви мають бути одного регістру). Вважаємо, що текст складається з $M = 33$ різних знаків - 33-ьох букв українського алфавіту. Закодуйте цей текст, скориставшись шифром зсуву/підстановки.

Для декодування тексту використайте ПММ $\lambda = (\mu, A, B)$, де:

- $T = 50000$ - кількість спостережень (об'єм зашифрованого тексту);
- $\{(X_n)\}_{n\geq 0}$ прихований ланцюг Маркова з $N = 33$ станами, матрицею перехідних ймовірностей A та початковим розподілом $\mu = (\mu_1, \ldots, \mu_N)$.
  Матрицю A слід побудувати наступним чином. Взяти текст (100 000 знаків чи й більше), подібний до зашифрованого за жанром, стилем або це може бути текст того ж автора, що й зашифрований текст. Визначте A_{ij} як кількість спостережень впорядкованої пари букв (i, j) в цьому тексті (а=1,б=2,...,я=33). Щоб всі елементи матриці А були невід'ємними і щоб матриця була стохастичною, додамо до кожного A_{ij} число 5, а потім нормуємо матрицю, поділивши кожен елемент матриці на суму елементів відповідної стрічки. В якості вектора μ візьмемо стаціонарнй розподіл, що відповідає побудованій матриці А, який визначається як розв'язок рівняння $\mu A = \mu$.
- матриця B містить умовні ймовірності $B_{ij} = P(Y_n = j | X_n = i)$, $i = \overline{1, N}, j = \overline{1, M}$.

Розв'яжіть спочатку задачу навчання для ПММ, використавши 1000 перших знаків зашифрованого тексту і 200 ітерацій алоритму Баума-Велша. В процесі навчання переоцінювати матрицю А та вектор μ не потрібно. В якості першого наближеня матриці B покладіть $B_{ij} \sim 1/33$.

Проаналізуйте отриману в результаті навчання матрицю В для отримання ключа шифру ($arg\max_j B_{ij}$ задає найбільш ймовірний відповідник знаку i оригінального тексту у зашифрованому). Якою виявилася частка правильно декодованих знаків?

Отримавши модель $\lambda = (\mu, A, B)$,застосуйте алгоритм Вітербі для декодувння всього тексту. Чи вдалося прочитати текст?