Марчук Г. В., старший викладач кафедри КН,
Мисливий М. В., студент групи ПІ-59
Державний університет «Житомирська політехніка»

ВИКОРИСТАННЯ ВКЛАДЕНЬ СЛІВ ТА МЕТОДУ K-НАЙБЛИЖЧИХ СУСІДІВ ДЛЯ ВИЗНАЧЕННЯ CEFR РІВНЯ СЛОВА

Common European Framework of Reference for Languages (CEFR) розшифровується як загальноєвропейські компетенції володіння іноземною мовою і є загальновживаним стандартом визначення рівня знання мови учня. CEFR визначає знання мови беручи до уваги вміння вивчаючого читати, слухати, говорити та писати нею. Всіх хто вивчає мову розподіляються на учнів початкового, середнього і високого рівня та за стандартом маркуються як A1, A2, B1, B2, C1, C2 (Рис.1).

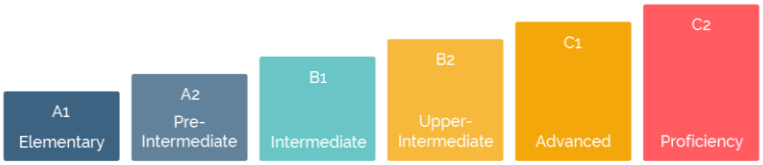

Рис. 1. Рівні знання англійської мови CEFR

Маючи CEFR стандарт можна використати його як основу для визначення рівня складності будь-якого слова. Визначимо слово рівня N, де N – одне з можливих значень рівня мови за стандартом CEFR, як слово, що повинно бути відоме вивчаючому мову, знання якого оцінено за стандартом CEFR на рівень N.

Векторне подання слів (англ. word embedding) – загальна назва низки підходів до мовного моделювання та навчання представлень в обробці природної мови, в яких слова або фрази зі словника відображають у вектори дійсних чисел. Концептуально воно дає математичне вкладення з простору з багатьма вимірами, по одному на слово, до неперервного векторного простору набагато меншої розмірності.

Так як слів багато, то ручне визначення рівня кожного слова стає дуже трудомістким заняттям, що потребує масу людино-годин. Для рішення проблеми можна використовуючи метод K-найближчих сусідів разом з вкладенням слів та невелику вибірку вже розмічених слів мови, що використовується.

Перед визначення рівня потрібного слова необхідно отримати векторні представлення слів з відомим CEFR рівнем використовуючи вкладення слів fastText (https://fasttext.cc/). На виході буде N векторів, після чого необхідно отримати векторне представлення слова з невідомим CEFR рівнем. Після чого застосувати алгоритм класифікації.

Існує безліч алгоритмів класифікації, але простота та високий рівень ефективності KNN (k-nearest neighbors algorithm) роблять його дуже популярним. Він також дуже гнучкий і може застосовуватися до будь-якого набору даних у формі вектору, не припускаючи нічого про його структуру чи походження (Рис.2).

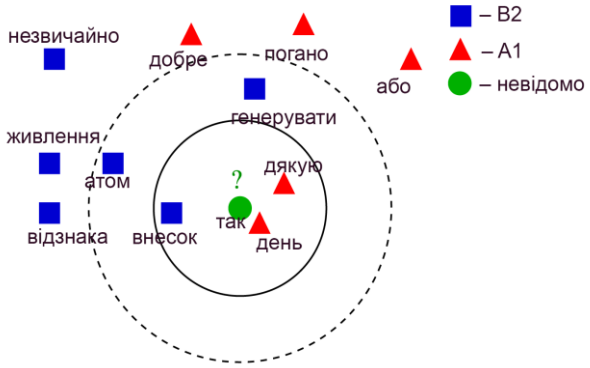

Рис. 2. Ілюстрація до методу визначення CEFR рівня слова

Після застосування методу K-найближчих сусідів, де найближчі сусіди це раніше отримані N векторів. На виході буде отримано класифіковане за CEFR рівнем слово.

Список використаних джерел

1. Enriching Word Vectors with Subword Information P. Bojanowski, E. Grave, A. Joulin, T. Mikolov.
2. Bag of Tricks for Efficient Text Classification A. Joulin, E. Grave, P. Bojanowski, T. Mikolov.
3. FastText.zip: Compressing text classification models A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jegou, T. Mikolov.
4. Cunningham, P., & Delany, S. J. (2021). k-Nearest Neighbour Classifiers - A Tutorial. ACM Computing Surveys, 54(6), 1–25. doi:10.1145/3459665.