АНОТАЦІЯ

Дипломну роботу виконано на 55 аркушах, вона містить 2 додатки та перелік посилань на використані джерела з 14 найменувань. У роботі наведено 5 рисунки та 2 таблиці.

Метою даної дипломної роботи є створення системи аналізу даних авіакатастроф для виявлення та представлення їх причин у великих масивах даних за 1919 – 2018 рр.

У роботі проведено аналіз існуючих програмних рішень для обробки текстової інформації та методів кластеризації - ієрархічна кластеризація, кластеризація k-середніх, expectation-maximization метод (EM - метод) та метод DBSCAN. Виконано їх порівняння з погляду на їх швидкість, ефективність, тип моделі, можливість обробки великої кількості даних та складність алгоритмів. На основі сформульованих критеріїв для розв’язання поставленої задачі обрано метод кластеризації k-середніх. Для перетворення текстових описів у вектори було обрано метод Bag-of-words з використанням міри TF-IDF.

Розроблено автоматизовану систему, що реалізує обрані методи. Виконано тестування розробленої системи.

Ключові слова: текстові описи авіакатастроф, векторизація тексту, корпус документів, кластеризація, анотація кластерів.