**Міністерство освіти і науки України**

**Львівський національний університет імені Івана Франка**

**Кафедра системного проектування**

“ЗАТВЕРДЖУЮ”

В.о. декана факультету електроніки та комп'ютерних технологій

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Половинко І.І. “\_\_\_\_\_\_”\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2018 року

**Опрацювання даних**

**ПРОГРАМА**

**навчальної дисципліни підготовки бакалаврів спеціальності**

**122 Комп’ютерні науки**

**факультету електроніки та комп'ютерних технологій**

**Львів 2018**

РОЗРОБНИК ПРОГРАМИ:

Демків Л.С. канд. фіз.-мат. наук, доцент кафедри системного проектування

Затверджено на засіданні Вченої ради факультету електроніки та комп'ютерних технологій “\_\_\_\_\_” \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2018 року, протокол №\_\_\_

 Голова Вченої ради Половинко І.І.

Обговорено та рекомендовано до затвердження Навчально-методичною радою факультету електроніки та комп'ютерних технологій

“\_\_\_\_\_” \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2018 року, протокол №\_\_\_

Голова Навчально-методичної ради

факультету електроніки та комп'ютерних технологій Шувар Р.Я.

Робочу програму схвалено на засіданні кафедри системного проектування

“\_\_\_\_”\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2018 року, протокол № \_\_\_

Завідувач кафедри системного проектування Шувар Р.Я.

# ВСТУП

Програма вивчення навчальної дисципліни “ **Опрацювання даних**” складена відповідно до освітньо-професійної програми підготовки бакалаврів напряму 126 Інформаційні системи та технології .

**Предметом** вивчення навчальної дисципліни є вивчення основних підходів до попередньої обробки (preprocessing) реальних даних для їх оптимального використання в алгоритмах аналізу даних. До методів попередньої обробки відносять: очищення даних, нормалізація даних, приведення типів даних, використання сучасних типів даних, трансформація даних, регресія даних, зменшення кількості даних, аналіз статистики даних. Дисципліна допомагає студентам набути необхідних практичних вмінь для роботи з неідеальними реальними даними в сучасних середовищах розробки.

**Міждисциплінарні зв’язки**: Для вивчення дисципліни необхідні знання з таких дисциплін: вища математика, теорія алгоритмів, алгоритмізація і програмування, об’єктно-орієнтоване програмування. Знання отримані під час вивчення дисципліни будуть використовуватися при вивченні наступних дисциплін: аналіз даних, машинне навчання.

**Програма навчальної дисципліни** складається з таких модулів:

***Модуль 1.***

 - Типи і підходи до збору даних

- Типи даних

- Кількісна і якісна оцінка даних структури даних (списки, вектори, набори, хеш-таблиці) та сховища даних

- графи та структури даних на графах

- OLAP-куби

***Модуль 2.***

- форматування і представлення даних

- фільтрування даних

- редукція даних

- поняття навчальної та тестової вибірки, їхнє створення та використання у різних моделях

навчання

**1. Мета та завдання навчальної дисципліни**

1.1. Метою викладання навчальної дисципліни “**Опрацювання даних**” є формування у студентів навиків та вмінь необхідних для отримання та створення даних, подання їх в в графічній формі, вивчення методів і алгоритмів обробки даних, визначення статистичних параметрів даних.

1.2. Основними завданнями вивчення дисципліни є забезпечити знайомство студентів з особливостями обробки даних, вивчення алгоритмів попередньої обробки даних; опанування теоретичного матеріалу і практичне оволодіння сучасними графічно-інформаційними технологіями, комп'ютерними та програмними засобами створення та опрацювання даних.

1.3. Згідно з вимогами освітньо-професійної програми студенти повинні

**Знати**: принципи формування та збереження даних; способи візуалізації даних; методи побудови, перетворення і аналізу графічної інформації; технології представлення даних в різних типах змінних,

**Вміти:** використовувати сучасні програмні засоби для попередньої обробки неідеальних реальних даних, для запису даних у відповідні структури даних.

.

**2.Інформаційний обсяг навчальної дисципліни**

1. Поняття даних. Визначення даних. Схеми даних. Основні особливості і недоліки сучасних реальних даних. Формати зберігання даних. Джерела даних: файли, бази даних, web-сервіси. Зчитування та запис файлових даних. Методи для роботи з потоками даних.
2. Опрацювання даних представлених класом. Об’єкти. Колекції об’єктів. Робота з даними в текстових форматах JSON XML/RDF TXT/CSV/Markdown. структури даних (списки, вектори, набори, хеш-таблиці) та сховища даних - графи та структури даних на графах - OLAP-куби
3. Використання методів лінійної алгебри при роботі з даними представленими векторами та матрицями. SVD розклад матриць та його використання при наданні рекомендацій.
4. Статистичний аналіз даних. Статистичні моменти даних. Роботи з великою кількістю даних (обновлення статистики даних)
5. Нормальний розподіл даних. Види розподілів.
6. Коваріація та кореляція даних. Регресія даних.
7. Опрацювання текстових даних.
8. Метод головних компонент.
9. Методи представлення даних, як початкових, так і отриманих в результаті обробки, з

метою візуальної оцінки їхньої якості чи інформативності, а також з метою демонстрації їхньої цінності для кінцевого споживача.

- Візуалізація даних як етап аналізу даних

- підходи до візуалізації даних та методи візуалізації даних (табличний, графічний, ієрархічний, картографічний, об’ємні візуальні моделі)

- засоби візуалізації даних (Tableau, ggplot2, gnuplot, matplotlib)

- наукова візуалізація (візуалізація тензорних полів, методи світлових променів. )

- редукція даних для візуалізації (Data Visualization також передбачає зменшення розмірності ознак досліджуваного об’єкта (наприклад перехід від 3-х вимірного простору ознак до 2-х вимірного та аналіз вибору ознак які є вагомими та мають якийсь стан досліджуваного об’єкта))

1. - Багатовимірне шкалювання
2. Програмні пакети для Data Cleansing/Transforming:

- OpenRefine (Opensource) є автономним desktop app з відкритим вихідним кодом для очищення даних і перетворення в інші формати. Він схожий на програми електронних таблиць (і може працювати з форматами файлів електронних таблиць), однак він веде себе як база даних.- Wrangler (Opensource) – Academic Data Cleansing

- Nifi (Opensource) - це програмний проект від Apache Software Foundation, призначений для автоматизації потоку даних між програмними системами. Він базується на програмному забезпеченні "NiagaraFiles", розробленому раніше NSA.

- SQOOP (Opensource) - набір інструментів для ефективного масового імпорту та експорту даних між системами структурованого зберігання даних, таких як реляційні БД, і NoSQL системами або промисловими warehouse-сховищами великих масивів даних, наприклад, Apache Hive, HBase або HDFS, що працюють на базі платформи Apache Hadoop. З реляційних СУБД підтримуються MySQL, PostgreSQL, Oracle, SQL Server і IBM DB2.

- Python + NumPy

- Python + Pandas

***3. Рекомендована література***

1. . Han, Jiawei. Data mining : concepts and techniques / Jiawei Han, Micheline Kamber, Jian Pei. – 3rd ed. ISBN 978-0-12-381479-1 Chapter 3. Data preprocessing
2. Michael R. Brzustowicz Data Science with Java
3. Методичні вказівки до виконання лабораторних робіт з дисципліни “Інженерія прикладних інтелектуально-орієнтованих програмних продуктів” для студентів спеціальностей 121 “Інженерія програмного забезпечення” та 122 “Комп’ютерні науки та інформаційні технології” (всіх форм навчання) / В.М. Льовкін. – Запоріжжя : ЗНТУ, 2016. – 80 с.

**4. Форма підсумкового контролю успішності навчання**

Підсумкову оцінку якості засвоєння навчальної програми з дисципліни “Опрацювання даних” визначають за результатами заліку, порядок проведення якого встановлює робоча навчальна програма.

Рівень засвоєння навчального матеріалу дисципліни визначають, використовуючи рейтингову систему оцінювання. Положення про рейтингову систему оцінювання знань розробляють та затверджується на засіданні кафедри з урахуванням особливостей професійної підготовки та розподілу навчального часу за видами занять. Це положення входить до складу робочої навчальної програми.

**5.Засоби діагностики успішності навчання**

Оцінка якості засвоєння навчальної програми включає поточний контроль успішності, два модульних контролі та складання заліку. Для поточного контролю засвоєння студентами навчального матеріалу передбачається виконання та захист лабораторних робіт, перелік яких наводиться в робочій навчальній програмі.

Для модульних контролів засвоєння студентами навчального матеріалу передбачається виконання двох модульних тестових робіт, порядок проведення та зміст яких наводяться в робочій навчальній програмі.

Організація індивідуальної роботи студентів визначається робочою навчальною програмою.