МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЗАПОРІЗЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ФАКУЛЬТЕТ МАТЕМАТИЧНИЙ
КАФЕДРА КОМП'ЮТЕРНИХ НАУК

ЗАТВЕРДЖУЮ
Декан математичного факультету
_____ __С.І. Гоменюк _
(підпис) (ініціали та прізвище)

«_____»_____

ВЕЛИКІ ДАНІ. НАУКА ПРО ДАНІ

РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

підготовки магістра
спеціальності 122 – Комп'ютерні науки
освітньо-професійна програма – Комп'ютерні науки

Укладач : Кеберле Н.Г., доцент, к.т.н.

Обговорено та ухвалено
на засіданні кафедри комп'ютерних наук

Протокол №__1__ від "_23_"__08__2018 р.

Завідувач кафедри
_____ __С.Ю. Борю_____
(підпис) (ініціали, прізвище)

Ухвалено науково-методичною радою
математичного факультету

Протокол №_3_ від "_25_"__10__2018 р.
Голова науково-методичної ради
факультету
_____ ____О.С. Пшенична_____
(підпис) (ініціали, прізвище)

2018 рік

1. Опис навчальної дисципліни

<table>
<tr><td rowspan="2">Найменування показників</td><td rowspan="2">Галузь знань, спеціальність, освітня програма, рівень вищої освіти</td><td colspan="2">Характеристика навчальної дисципліни</td></tr>
<tr><td>денна форма навчання</td><td>заочна форма навчання</td></tr>
<tr><td rowspan="2">Кількість кредитів – 4</td><td>Галузь знань
12 – Інформаційні технології</td><td colspan="2">нормативна</td></tr>
<tr><td></td><td colspan="2">цикл професійної підготовки</td></tr>
<tr><td rowspan="2">Розділів – 2</td><td rowspan="2">Спеціальність
122 – Комп’ютерні науки</td><td colspan="2">Рік підготовки:</td></tr>
<tr><td rowspan="2">Загальна кількість годин – 120</td><td rowspan="2">2-й</td><td rowspan="2">2-й</td></tr>
<tr><td>Освітньо-професійна програма
Комп’ютерні науки</td></tr>
<tr><td rowspan="8">Тижневих годин для денної форми навчання:
аудиторних – 3 год.
самостійної роботи студента – 8 год.</td><td rowspan="8">Рівень вищої освіти:
магістерський</td><td colspan="2">Лекції</td></tr>
<tr><td>12 год.</td><td>4 год.</td></tr>
<tr><td colspan="2">Лабораторні</td></tr>
<tr><td>22 год.</td><td>8 год.</td></tr>
<tr><td colspan="2">Самостійна робота</td></tr>
<tr><td>86 год.</td><td>108 год.</td></tr>
<tr><td colspan="2">Вид контролю:</td></tr>
<tr><td>залік</td><td>залік</td></tr>
</table>

2. Мета та завдання навчальної дисципліни

Метою викладання навчальної дисципліни «Великі дані. Наука про дані» є ознайомлення з галуззю великих даних (Big Data), пов'язаних даних (Linked Data) та з інструментарієм, що використовується для створення, зберігання, публікації, пошуку, аналізу великих даних.

Основними **завданнями** вивчення навчальної дисципліни «Великі дані. Наука про дані» є: оволодіння основними поняттями в галузі великих даних, ознайомлення з принципами організації сховищ великих даних, із базовими алгоритмами збереження та пошуку у сховищі великих даних, ознайомлення із переліком задач, що відносяться до організації роботи з великими даними, та способами рішення таких задач, формування навичок з реалізації сховища великих даних, організації процесів роботи зі сховищем, типового аналізу великих даних.

У результаті вивчення навчальної дисципліни студент **повинен**

знати:

- сутність понять "великі дані", "пов'язані дані";
- властивості, засоби опису, публікації, використання, зберігання пов'язаних даних;
- мови запитів до пов'язаних даних;
- властивості, особливості та джерела великих даних;
- принципи класифікації великих даних;
- основні архітектури сховищ великих даних;
- основні алгоритми роботи зі сховищем великих даних (MapReduce);
- основні задачі аналізу у постановці для великих даних;
- основні задачі, в яких застосовується інтелектуальний аналіз даних.

вміти:

- розгортати точку доступу до пов'язаних даних та створювати запити до них;
- створювати додатки, що використовують пов'язані дані;
- публікувати пов'язані дані;

- розгортати Apache Spark інфраструктуру;
- створювати додатки, що використовують великі дані;
- застосовувати інструменти Spark для аналітичної обробки великих даних.

Згідно з вимогами освітньо-професійної програми студенти повинні досягти таких **результатів навчання (компетентностей)**:

- здатність до абстрактного мислення, аналізу та синтезу;
- здатність до пошуку, оброблення та аналізу інформації з різних джерел;
- здатність до математичного та логічного мислення, формулювання та досліджування математичних моделей, обґрунтовування вибору методів і підходів для розв'язування теоретичних і прикладних задач в галузі комп'ютерних наук, інтерпретування отриманих результатів;
- здатність реалізувати багаторівневу обчислювальну модель на основі архітектури клієнт-сервер, включаючи бази даних, для забезпечення обчислювальних потреб багатьох користувачів, обробки транзакцій, у тому числі на хмарних сервісах.

Міждисциплінарні зв'язки:

Вивчення дисципліни «Великі дані. Наука про дані» базується на знаннях, отриманих під час вивчення дисциплін «Алгоритми і структури даних», «Дискретна математика (для програмістів)», «Процедурне програмування», «Бази даних та інформаційні системи», «Інформаційні мережі».

Знання та уміння, отримані під час вивчення дисципліни «Великі дані. Наука про дані», можуть бути використані у кваліфікаційних роботах магістрів.

3. Програма навчальної дисципліни

Розділ 1. Робота з Linked Open Data – відкритими пов'язаними даними.

Тема 1. *Принципи організації мережі пов'язаних даних.*

Визначення пов'язаних даних. Модель посилання у мережі пов'язаних даних - RDF, модель посилання у мережі пов'язаних даних із частковим структуруванням ресурсів - RDF Schema. Точки доступу до пов'язаних даних. Принципи організації мережі пов'язаних даних. Каталоги наборів пов'язаних даних (DCAT, тематичні набори). Онтології для анотування пов'язаних даних. Каталоги онтологій (LOV). Принципи «5 star» для пов'язаних даних. Застосування пов'язаних даних.

Тема 2. *Створення пов'язаних даних та аналіз пов'язаних даних у аналітичних платформах.*

Принципи створення пов'язаних даних. Структура IRI/URI. Резолюція IRI/URI. Платформи для конвертації звичайних даних у пов'язані (OpenRefine, RDB2RDF, GraphDB).

Розділ 2. Методи роботи з Big Data.

Тема 3. *Принципи організації сховищ великих даних.*

Визначення великих даних. Характеристики великих даних: 5"V" – volume, velocity, variety, value, veracity. Джерела великих даних – внутрішньоорганізаційні та зовнішні: соціальні мережі, дані сотових операторів, журнали подій у мережі комп'ютерів, історії замовлень у онлайн-магазинах, історії переглянутих сторінок, дані про використання пластикових карток, та інші.

Збереження великих даних. Технології: in-memory (Oracle Exadata, SAP HANA), NoSQL, Hadoop/MapReduce. Алгоритм MapReduce.

Hadoop – інфраструктура для збереження та обробки великих даних. Компоненти Hadoop – розподілена файлова система HDFS, метод розподіленого виконання програм MapReduce, система управлення ресурсами кластерів YARN.

Розгортання систем обробки великих даних у платформах хмарних обчислень Google Cloud.

Тема 4. *Інструменти аналітичної обробки даних у сховищі великих даних.*

Apache Spark – інфраструктура кластерних обчислень. Компоненти Spark – MapReduce, HDFS, YARN, Hive, Pig, Zookeeper, Flume та інші. Алгоритм Apache Hive. Мова Pig Latin (Apache Pig). Планування завдань Apache Oozie.

4. Структура навчальної дисципліни

Назви розділів і тем	Кількість годин							
	Денна форма				Заочна форма			
	усього	у тому числі			усього	у тому числі		
		л	лаб	сам. роб.		л	лаб	сам. роб.
1	2	3	5	6	7	8	10	11
Розділ 1. Робота з Linked Open Data – відкритими пов'язаними даними.								
Тема 1. Принципи організації мережі пов'язаних даних.	21	2	4	15	28	1	2	25
Тема 2. Створення пов'язаних даних та аналіз пов'язаних даних у аналітичних платформах	35	2	8	25	31	1	2	28
Разом за розділом 1	56	4	12	40	59	2	4	53
Розділ 2. Методи роботи з Big Data.								
Тема 3. Принципи організації сховищ великих даних.	28	4	4	20	30	1	2	27
Тема 4. Інструменти аналітичної обробки даних у сховищі великих даних.	36	4	6	26	31	1	2	28
Разом за розділом 2	64	8	10	46	61	2	4	55
Усього годин	120	12	22	86	120	4	8	108

5. Теми лекційних занять

№ з/п	Назва теми	Кількість годин	
		денна форма	заочна форма
1	Linked Data. Причини появи. Методи обробки. SPARQL сховища триплів (триплстори). Federated SPARQL.	2	1
2	Створення пов'язаних даних та аналіз пов'язаних даних у аналітичних платформах.	2	1
3	Визначення "наука про дані". Властивості великих даних (Big Data). Джерела великих даних та основні шляхи використання.	4	1
4	Інструменти пошуку та аналітичної обробки даних у сховищі великих даних.	4	1
Разом:		**12**	**4**

6. Теми лабораторних занять

№ з/п	Назва теми	Кількість годин	
		денна форма	заочна форма
1	Виконання SPARQL-запитів з візуалізацією у HTML. Знайомство з бібліотекою Jena API. Програмування простого додатку виконання SPARQL-запитів за допомогою Jena API.	4	2
2	Аналіз даних в RapidMiner. Створення процесу завантаження та пошуку моди/медіани у наборі даних.	8	2
3	Знайомство зі стеком технологій Apache Spark.	4	2
4	Інструменти аналітичної обробки даних у Google Cloud Platform.	6	2
Разом:		**22**	**8**

7. Теми самостійної роботи

№ з/п	Назва теми	Кількість годин	
		денна форма	заочна форма
1	Встановлення та налагодження графової семантичної бази даних (сховища триплів) Ontotext GraphDB	15	25
2	Опрацювання набору даних у форматі CSV для перетворення у RDF за допомогою Ontotext GraphDB OpenRefine.	12	20
2	Встановлення RapidMiner та його налагодження для опрацювання RDF даних.	13	8
3	Реєстрація у Google Cloud Platform. Знайомство з можливостями платформи GCP.	20	27
4	Застосування GCP BigQueryML для аналітичних задач	26	28
Разом:		**86**	**108**

8. Види контролю і система накопичення балів

Видами контролю з дисципліни «Великі дані. Наука про дані» є:

- *перевірка знань: опитування після виконання самостійної роботи;*
- *перевірка вмінь: виконання та захист лабораторних робіт;*

Система накопичення балів

Поточний контроль знань		Підсумковий контроль	Сума
		залік	
Розділ 1	Розділ 2	40	100
30	30		

№	**Вид контрольного заходу**	**Кількість**	**Кількість**	**Усього**

		контрольних заходів	балів за 1 захід	балів
1	Виконання самостійних робіт 1-3	3	6	18
2	Підготовка та захист лабораторної роботи 1	1	7	7
3	Підготовка та захист лабораторної роботи 2	1	5	5
4	Виконання самостійних робіт 4-5	2	5	10
5	Підготовка та захист лабораторних робіт 3-4	2	10	20
7	Залік. Контрольне тестування за результатами вивчення матеріалу Розділів 1, 2	1	40	40
	Усього	10		100

Шкала оцінювання: національна та ECTS

За шкалою ECTS	За шкалою університету	За национальною шкалою	
		Екзамен	Залік
A	90 – 100 (відмінно)	5 (відмінно)	Зараховано
B	85 – 89 (дуже добре)	4 (добре)	
C	75 – 84(добре)		
D	70 – 74 (задовільно)	3 (задовільно)	
E	60 – 69 (достатньо)		
FX	35 – 59 (незадовільно – з можливістю повторного складання)	2 (незадовільно)	Не зараховано
F	1 – 34 (незадовільно – з обов'язковим повторним курсом)		

9. Рекомендована література

ОСНОВНА

1. Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. Москва : Издательство «Манн, Иванов и Фербер», 2014. 240 с.
2. Силен Д., Мейсман А., Али М. Основы Data Science и Big Data: Python и наука о данных. Санкт-Петербург : ПИТЕР, 2017. 336 с.
3. RapidMiner 4.2 User Guide Operator Reference Developer Tutorial. URL : http://moodle.znu.edu.ua/pluginfile.php?file=/89874/mod_resource/content/1/15967723-Rapidminer-4-2-Tutorial.pdf (дата звернення 20.08.2018)
4. Дьяконов А.Г. Анализ данных, обучение по прецедентам, логические игры, системы WEKA, RapidMiner и MatLab (Практикум на ЭВМ кафедры математических методов прогнозирования): учебное пособие. Москва : Издательский отдел факультета ВМК МГУ имени М.В. Ломоносова, 2010. 278 с.

ДОДАТКОВА

1. Zikopoulos P., deRoos D., Bienko C., Buglio R., Andrews M. Big Data Beyond the Hype. A Guide to Conversations for Today's Data Center. McGrawHill Education. 2015. 393 p.
2. Sathi A. Big Data Analytics. Disruptive Technologies for Changing the Game. MC Press Online LLC. 2013. 93 p.

10. Інформаційні ресурси:

1. GraphDB OntoText. URL: http://graphdb.ontotext.com/documentation/free/introduction-to-semantic-web.html
2. KDNuggets: Data Mining Community Top Resource for Analytics, Data Mining, and Data Science Software, Companies, Data, Jobs, Education, News, and more. URL: http://www.kdnuggets.com
3. The Data Mine. URL: http://www.the-data-mine.com
4. Google Cloud. https://cloud.google.com/

Доповнення та зміни до робочої програми навчальної дисципліни

Протокол засідання кафедри (дата та номер)	Внесені зміни	Підпис завідувача кафедри, дата