ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Змістовий модуль 1. **Поняття, методи та продукти комп'ютерної лінгвістики**

Тема 1. Вступ. Лінгвістика та її структура

Передмова. Мета курсу та цільова аудиторія. Зв'язки комп'ютерної лінгвістики з галузями інформатики та штучним інтелектом. Обробка природної мови. Лінгвістика та її структура. Поняття та межі галузі комп'ютерної лінгвістики. Складові комп'ютерної лінгвістики. Квантитативна лінгвістика. Корпусна лінгвістика. Статистична лінгвістика. Математична лінгвістика. Поняття слова. Використання підходів і методів фундаментальної науки. Сучасний стан прикладних досліджень з комп'ютерної лінгвістики.

Тема 2. Базові поняття лінгвістики

Загальні поняття про мову. Мова і мислення. Мова як знакова система. Мова і мовлення. Мовна структура. Лінгвістичні мови. Фонетика. Звуковий характер мови. Фонологія. Поняття про фонеми. Звуки і букви. Поняття про алфавіт. Морфологія. Основні поняття. Морфема. Види морфем. Формальні моделі морфології. Морфологічні словники. Синтаксис. Основні поняття. Словосполучення. Речення. Синтаксичні відносини. Керування. Узгодження. Примикання. Комунікативна організація речення. Семантика. Значення. Види значень. Семантичні відносини. Парадигматичні та синтагматичні відносини. Формальні моделі семантики. Основи лексикографії. Семантичні мережі. Фрейми. Лексико-семантичні комплекси. Прагматика. Основні поняття. Види прагматичних знань. Проблема розуміння текстів. Умови цільового розуміння текстів.

Тема 3. Розвиток ідей, підходів і методів комп'ютерної лінгвістики

Структурний підхід в комп'ютерній лінгвістиці. Внесок Хомського у комп'ютерну лінгвістику. Проста контекстно-вільна граматика. Трансформаційні граматики. Лінгвістичні дослідження після Хомського. Валентності та тлумачення. Труднощі теорії Хомського. Граматика структури фраз, керована головною складовою. Ідея уніфікації. Теорія ЗМІСТ ↔ ТЕКСТ: мова як багатоступеневий перетворювач. Моделі керування. Дерева залежностей. Семантичні зв'язки.

Тема 4. Огляд сучасних продуктів комп'ютерної лінгвістики

Класифікація прикладних лінгвістичних систем. Препроцесинг текстів. Парсинг, тегування, стеммінг і лемматизація. Автоматична розстановка переносів. Перевірка орфографії. Перевірка граматики. Перевірка стилістики. Пошук слів і словосполучень. Підсумування тематики документа. Видобування фактичних даних з текстів. Інформаційні пошукові системи. Автоматичний переклад. Інтерфейс природною мовою. Генерування текстів. Системи розуміння мови та пов'язані з ними системи.

Змістовий модуль 2. **Теорії та моделі в основі алгоритмів комп'ютерної лінгвістики**

Тема 5. Мова як двонапрямлений перетворювач змісту та тексту

Можливі точки зору стосовно природної мови. Мова як двонапрямлений перетворювач. Поняття тексту і змісту. Два способи представлення змісту. Розкладання і атомізація змісту. Неоднозначність "картування» ЗМІСТ ↔ ТЕКСТ: синоніми і омоніми. Подальша інформація про омонімію. Багатоетапність перетворення ЗМІСТ ↔ ТЕКСТ. Переклад як багатоступінчасте перетворення.

Тема 6. Лінгвістичні знаки

Два боки поняття знака. Лінгвістичний знак. Лінгвістичний знак у теорії ЗМІСТ ↔ ТЕКСТ. Лінгвістичний знак у теорії граматики структури фраз, керованої головною складовою. Змістовний знак: умовність чи дане природою? Порівняння генеративних ідей, ідей з теорії ЗМІСТ ↔ ТЕКСТ та ідеї обмежень.

Тема 7. Лінгвістичні моделі

Поняття моделювання. Нейролінгвістичні моделі. Психолінгвістичні моделі. Функціональні моделі мови. Дослідження лінгвістичних моделей. Спільні риси сучасних моделей мови. Особливості моделі ЗМІСТ ↔ ТЕКСТ. Зредуковані моделі. Практична потреба в лінгвістичних моделях. Анало-

гії в природних мовах. Емпіричні підходи проти раціоналістичних підходів. Обмеженість сфери застосування сучасних лінгвістичних теорій.

Змістовий модуль 3. **Статистична лінгвістика**

Тема 8. Тексти як складні системи та мережі

Поняття систем. Системи із взаємодіючими елементами. Складні системи. Складні мережі. Лінгвістичні компоненти як елементи складних систем. Комп'ютерна лінгвістика і теорія складних систем. Лінгвістичні мережі. Основні характеристики мереж.

Тема 9. Основні поняття лінгвістичної статистики та складних систем

Детермінізм і стохастичний підхід. Поняття нульової статистичної гіпотези. Статистичні підходи в лінгвістиці. Випадкові величини. Генеральна сукупність і вибірка. Репрезентативність лінгвістичної вибірки. Дискретні та неперервні випадкові величини. Розподіли ймовірності випадкових величин. Густина розподілу ймовірності. Кумулятивна функція розподілу. Рівномірний розподіл. Розподіл Пуасона. Нормальний розподіл. Правило «трьох сигм» і центральна гранична теорема. Експоненційний і розширений експоненційний розподіли. Розподіли Вейбуля та Ґуда. Степеневий розподіл. Поняття скейлінгу. Лог-нормальний розподіл. «Оманливі» степеневі розподіли. «Логарифмічна» рангова залежність.

Імовірність екстремальних подій. «Важкий» і «легкий» хвости розподілів. «Справедливі» та «несправедливі» розподіли. Статистичні моменти розподілів з важким хвостом. Середнє та дисперсія. Проблема існування моментів.

Приклади статистичних розподілів поза межами лінгвістики. Зріст людини. Населення міст. Сила землетрусів. Розподіл багатства між людьми. Відвідування сайтів. Імена та прізвища. Спорт. Популярність наукових журналів, наукових статей і авторів.

Тема 10. Методичні особливості вивчення лінгвістичної статистики

Шуми в розподілах імовірності. Бінування та побудова гістограм. Перехід до кумулятивного розподілу. Візуалізація та «лінеаризація» розподілів імовірності різних типів. Методика досліджень степеневих розподілів. Лінійна апроксимація в логарифмічному масштабі. Нелінійна апроксимація та її проблеми. Метод Ньютона. Логарифмічне зважування. Алгоритми визначення показника степеневих розподілів дискретних випадкових величин. Метод максимальної схожості. Тест «хі-квадрат» у статистичній лінгвістиці. Інші статистичні тести.

Корпуси текстів. Автоматичне завантаження з Інтернету. Особливості статистики для корпусної лінгвістики. Поняття ергодичності системи.

Тема 11. Основні статичні та динамічні закони лінгвістики

Слова, слововживання, словоформи та лемми. Довжина і словник тексту. Абсолютна та відносна частота слова. Рангові залежності частот слів. Алгоритмізація побудови частотних лексичних таблиць. Перший і другий закони Ціпфа для лексики. Поняття про лексичний спектр. Закон Парето. Закони статистичної лінгвістики комп'ютерних програм. Співвідношення загальної кількості елементів і кількості класів цих елементів. Лексичний словник. Закон Гіпса–Гердана. Синтетичні та аналітичні мови. Закони Гіпса для мов програмування та текстів програм. Метод ковзного вікна у вивченні словника. Закон Гіпса для корпусів текстів. Взаємні зв'язки експонент різних степеневих законів.

Тема 12. Закони статистики для інших лінгвістичних рівнів. Статистика інших складних систем

Альтернативи степеневим законам в лінгвістиці. Логарифмічні та експоненційні рангові залежності. Інші закони для розподілів імовірності та динаміки словника. Частота букв. Особливості статистики графем і фонем. Рангові та частотні залежності для букв. «Словник» літер і закон Гіпса для букв. Поняття про лінгвістичні n-грами. Лексичні та буквені (символьні) n-грами. Статистичні закони для n-грам і їхні зв'язки з законами для лексики та графем. Узагальнення закону Гіпса для лексичних n-грам. Закономірності для складів. Ієрогліфічні мови. Випадки китайської та корейської мов з обмеженим словником. Словник рандомних текстів мавпи Міллера з алфавітом на одну букву.

Розподіли ймовірності для імен і прізвищ. Прізвища в Кореї. Рангові залежності в спортивних турнірах. Популярність наукових журналів і цитування наукових статей. Популярність автора. Індекс Герша.

Тема 13. Інші закони лінгвістичної статистики

Частоти речень і слів різних довжин. Часова еволюція частоти літер. Рангові та частотні залежності для слів різних довжин. Середня довжина слова в різних мовах. Особливості флуктуацій частоти літер. «Шуми» в рангових залежностях для букв. Інформаційна ентропія Шеннона. Закони, що пов'язують лінгвістичні елементи різних рівнів. Закон Менцерата–Альтмана.

Тема 14. Нульові стохастичні гіпотези в статистичній лінгвістиці. Рандомні тексти

Рандомні тексти. Типи рандомних текстів. Тексти Хомського. Тексти «мавпи Міллера». Модель зростання Саймона. Модифікована модель Саймона з нелінійним зростанням словника. Тексти Саймона з пам'яттю. Моделі зростання на основі урни Полі.

Тексти на основі ланцюжків Маркова. Рандомізовані природні тексти. Локальна та глобальна рандомізація на різних лінгвістичних рівнях. Інвертування текстів. Властивості рандомних текстів. Проблеми розрізнення природних і рандомних текстів. Стиснення природних і рандомних текстів.

Тема 15. Механізми появи степеневих розподілів

Основні підходи в поясненні законів статистичної лінгвістики. Наближене відтворення законів Ціпфа, Парето та Гіпса. Проблема «взаємодій» елементів системи. Принцип «багатий стає багатшим» у складних системах. Модель «дискурс–оточення». Принцип «переважного приєднання» в складних мережах. Тригерування інновацій. «Фазовий перехід» комунікація–відсутність комунікації. Показник степеневого розподілу лексичної ймовірності для текстів і проблема психічних захворювань. Особливості лексичних спектрів для дітей.

Змістовий модуль 4. **Основи опрацювання природної мови**

Тема 16. Характеристика повторюваності в текстах

Основні означення. Алгоритм обчислень. Режими обчислень характеристики повторень. Програмна реалізація. Деякі результати для характеристики повторень в основному режимі. Пошук «самоплагіату» в текстах.

Тема 17. Класифікація та порівняння текстів

Категоризація та класифікація текстів. Індексація та абстрагування текстів. Лінгвістичні маркери. Стилістика та її статистичні ознаки. Проблеми визначення мови, встановлення авторства, жанру та плагіату. Подібність текстів і «відстань» між текстами. Різні дефініції «відстаней». «Відстані» між мовами. Визначення «відстаней» в умовах завад (типографських помилок). «Відстані» між ранговими залежностями і лексичними спектрами.

Підходи на основі векторного простору. Скалярний добуток текстів. Перевизначені скалярні добутки. Розпізнавання текстового плагіату. «Кусковий» плагіат.

Тема 18. Автоматичний аналіз настроїв

Типи аналізу настроїв. Переваги автоматичного аналізу настроїв. Способи візуалізації даних аналізу настроїв. Принципи роботи аналізаторів настроїв. Аналізатори настроїв, засновані на правилах. Повністю автоматичні аналізатори настроїв. Гібридні аналізатори настроїв. Основні проблеми аналізу настроїв. Застосування аналізу настроїв. Деякі програмні засоби машинного аналізу настроїв.

Тема 19. Пошук ключових слів

Функціональні та змістові слова. Поняття ключових слів і виразів (n-грам). Загальні поняття. Абсолютні та відносні методи. «Відносний» метод TF-IDF. Інформаційний пошук на основі ключових слів. Метод Бріна–Пейджа в інформаційному пошуку.

Частотні ознаки ключових слів. Пошук ключових слів на основі явища кластеризації. Закон рідкісних подій і розподіл Пуасона. Час очікування рідкісної події. Приклади. Аварії та технічні поламки, землетруси, серцеві ритми. Лінгвістичні приклади. Експоненційний розподіл для часів очікування рідкісних подій. Відхилення ймовірності часів очікування ключових слів і n-грам од експоненційного розподілу як основа абсолютного методу. «Розширена експонента» та степеневі розпо-

діли. Розподіл Вейбуля. Розподіли ймовірності першої появи слів і лексичних n-грам у тексті. Семантика тексту. Розрізнення природних і рандомних текстів на основі кластеризації. Ключові слова в текстах комп'ютерних програм. Поняття про мережеві методи пошуку ключових слів.

Тема 20. Флуктуації та скейлінг у лінгвістиці та інших складних системах

Поняття скейлінгу. Масштабні залежності в комп'ютерній лінгвістиці. Закон Гіпса як приклад масштабної залежності. Масштабні залежності середніх довжин слів і речень.

Флуктуації. Відносні флуктуації. Приклади зі статистичної механіки. Мікроскопічне, мезоскопічне і макроскопічне наближення. Самоусереднювані статистичні величини. Поняття кореляцій. Флуктуації за наявності кореляцій.

Флуктуації в лінгвістиці. Відображення тексту на числовий ряд. Тематичні «неоднорідності» текстів як «заморожені» флуктуації. Короткосяжні та довгосяжні кореляції в лінгвістиці. Закон Тейлора. Ресурси флуктуаційного аналізу: кореляції, ключові слова та семантика тексту. Флуктуації частоти слів у текстах програм. Часова мова в описі флуктуацій. Метод рандомних прогулянок. Типи дифузії. Нормальна та аномальна дифузія.

Підхід ковзного вікна в дослідженні флуктуацій. Аналогії з галуззю цифрової обробки зображень. Практична реалізація алгоритму біжучого вікна. Методи FA для встановлення кореляцій. Поняття тренду в часовому ряді. Метод DFA для встановлення кореляцій на фоні трендів. Флуктуаційний аналіз на корпусах текстів.

Тема 21. Складні мережі. Ефект тісного світу в мережах

Основні поняття теорії складних мереж. Деякі властивості складних мереж: перколяція, ефект тісного світу та механізм переважного приєднання. Приклади складних мереж та їхні властивості. Лінгвістичні мережі. Означення зв'язків вузлів. Ефект «тісного світу» в лінгвістичних мережах. Мережеві методи пошуку ключових слів текстів.

Змістовий модуль 5. **Аналіз, розпізнавання та синтез природної мови. Машинний переклад та комп'ютерна лексикографія**

Тема 22. Автоматичне введення звуків мови, аналіз та розпізнавання мови комп'ютером

Акустична будова звуків мови. Загальні принципи організації мовної комунікації. Будова і принципи роботи мовного тракту людини. Теорія сприйняття звуків мови людиною. Аналого-цифрове перетворення мовних сигналів. Проблеми автоматичного розпізнавання усного мовлення. Методи аналізу та розпізнавання усного мовлення. Первинна обробка мовних сигналів. Методи розпізнавання окремих звуків. Звуковий спектрограф. Спектрально-часовий аналіз мовлення. Бінарна селекція звукових елементів мови. Методика формантного аналізу мови. Розпізнавання мовлення із залученням лінгвістичної інформації. Автоматичне розпізнавання усного мовлення.

Тема 23. Синтез мови та мовні технології

Розвиток систем синтезу усного мовлення. Метод кодування сигналів. Фонетичний синтез мовлення за допомогою формантних синтезаторів. Синтез мовлення за методом предикативного кодування. Алгоритм синтезу усного мовлення за текстом. Загальна структура синтезатора. Лінгвістичний, просодичний, фонетичний, акустичний та компіляційний синтезатори. Сучасні технології усного мовлення.

Тема 24. Машинний переклад і комп'ютерна лексикографія

Загальні положення машинного перекладу. Оцінка та критерії якості перекладу. Типові помилки машинного перекладу. Аналіз теоретичних моделей. Структура систем перекладу. Алгоритми аналізу та синтезу. Види перекладу та характеристики систем перекладу. Сучасний стан машинного перекладу. Предмет, методи та теорія комп'ютерної лексикографії. Проблеми комп'ютерної лексикографії.

Література

Основна:

1. Bolshakov I. Computational linguistics. Models, resources, applications / I. Bolshakov, A. Gelbukh. – Mexico : Ciencia de la Computacion, 2004. – 198 p.

2. Bird S. Natural language processing with Python / S. Bird, E. Klein, E. Loper. – Sebastopol : O'Reilly. – 2009. – 504 p.
3. Manning C. D. Foundations of statistical natural language processing / Manning C. D., Schutze H. – London : The MIT Press Cambridge, 1999. – 680 p.
4. Кушнір О. С. Основи комп'ютерної лінгвістики (конспект лекцій) / О. С. Кушнір. – Львів : Видавн. Львів. ун-ту, 2023. – 292 с.
5. Jurafsky D. Speech and language processing / D. Jurafsky, J. H. Martin. – New Jersey : Prentice Hall, 2023. – 628 p.
6. Clark A. The handbook of computational linguistics and natural language processing / A. Clark, C. Fox, S. Lappin. – Chichester : John Wiley & Sons, 2010. – 801 p.
7. Hausser R. Foundations of computational linguistics: Man–machine communication in natural language / R. Hausser. – Berlin : Springer, 1999. – 468 p.
8. Kracht M. Introduction to probability theory and statistics for linguistics / M. Kracht. – Oakland : UCLA, 2005. – 137 p.
9. Delmonte R. Computational linguistic text processing / New York : Nova Science Publishers, 2009. – 382 p.
10. Kornai A. Mathematical linguistics / A. Kornai. – London : Springer, 2007. – 300 p.
11. Web information retrieval / S. Ceri, A. Bozzon, M. Brambilla, E. Della Valle, P. Fraternali, S. Quarteroni. – Berlin : Springer, 2013. – 287 p.
12. de Araújo L. C. Statistical analyses in language usage / L. C. de Araújo. – Belo Horizonte : Universidade Federal de Minas Gerais, 2013. – 199 p.
13. Математична лінґвістика. Книга 1. Квантитативна лінгвістика / В. В. Пасічник, Ю. М. Щербина, В. А. Висоцька, Т. В. Шестакевич. – Львів : Новий світ – 2000, 2012. – 359 с.
14. Волошин В. Г. Комп'ютерна лінгвістика / В. Г. Волошин. – Суми : Університетська книга, 2004. – 382 с.
15. Мирам Г. Алгоритмы перевода: Вступительный курс по формализации перевода / Г. Мирам. – Киев : Эльга, Ника-Центр, 2004. – 176 с.
16. Хархалис Р. И. Компьютерный перевод иностранных текстов / Р. И. Хархалис. – Киев : Терези, 1998. – 193 с.

Додаткова:

17. Zanette D. H. Statistical patterns in written language / Zanette D. H. – Centro Atomico Bariloche, 2012. – 87 p. URL: http://fisica.cab.cnea.gov.ar/estadistica/2te/
18. Складні мережі // Ю. Головач, О. Олємской, К. фон Фербер, Т. Головач, О. Мриглод, І. Олємской, В. Пальчиков // Журн. фіз. дослідж. – 2006. – Т. 10, №4. – С. 247–289.
19. Newman M. E. J. Power laws, Pareto distributions and Zipf's law / Newman M. E. J. // Contemporary Phys. – 2005. – Vol. 46. – P. 323–351.
20. Ferrer i Cancho R. Zipf's law from a communicative phase transition / R. Ferrer i Cancho // Eur. Phys. J.: B. – 2005. – Vol. 47. – P. 449–457.
21. Kornai A. How many words are there? / A. Kornai // Glottometrics. – 2002. – Vol. 4. – P. 60–85.
22. Pilgrim C. Bias in Zipf's law estimators / C. Pilgrim, T. T. Hills // Sci. Rep. – 2021. – Vol. 11. – 17309 (12 pp.).
23. Espitia D. Universal and non-universal text statistics: Clustering coefficient for language identification / D. Espitia, H. L. Ridaura // Physica A. – 2020. – Vol. 553. – 123905 (25 pp.).
24. Simon H. On a class of skew distribution functions / H. Simon // Biometrika. – 1955. – Vol. 42. – P. 425–440.
25. Zanette D. H. Dynamics of text generation with realistic Zipf distribution / D. H. Zanette, M. A. Montemurro // J. Quant. Linguist. – 2005. – Vol. 12. – P. 29–40.
26. Cattuto C. A Yule-Simon process with memory / C. Cattuto, V. Loreto, V. D. P. Servedio // Europhys. Lett. – 2006. – Vol. 76. – P. 208–214.

27. Altmann E. G. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words / E. G. Altmann, J. B. Pierrehumbert, A. E. Motter // PLOS ONE. – 2009. – Vol. 4. – e7678 (7 pp.).
28. Altmann E. G. On the origin of long-range correlations in texts / E. G. Altmann, G. Cristadoro, M. D. Esposti // Proc. Nat. Acad. Sci. (USA). – 2012. – Vol. 109. – P. 11582–11587.
29. Ebeling W. Long-range correlations between letters and sentences in texts / W. Ebeling, A. Neiman // Physica A. – 1995. – Vol. 215. – P. 233–241.
30. Barabási A.-L. The origin of bursts and heavy tails in human dynamics / A.-L. Barabási // Nature. – 2005. – Vol. 435. – P. 207–211.
31. Goh K.-I. Burstiness and memory in complex systems / K.-I. Goh, A.-L. Barabási // Europhys. Lett. – 2008. – Vol. 81. – 48002 (5 pp.).
32. Brin S. The anatomy of a large-scale hypertextual Web search engine / S. Brin, L. Page. – Computer Networks and ISDN Systems. – 1998. – Vol. 30. – P. 107– 117.
33. The PageRank citation ranking: bringing order to the web / S. Brin, L. Page, R. Motwani, T. Winograd. – Technical Report. – Stanford : Stanford InfoLab, 1999. – 17 p.
34. Luhn H. P. The automatic creation of literature abstracts / H. P. Luhn // IBM J. Res. Development. – 1958. – Vol. 2, No 2. – P. 159–165.
35. Keyword detection in natural languages and DNA / M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz, A. M. Somoza // Europhys. Lett. – 2002. – Vol. 57. – P. 759–764.
36. Herrera J. P. Statistical keyword detection in literary corpora / J. P. Herrera, P. A. Pury // Eur. Phys. J. – 2008. – Vol. 63. – P. 135–146.
37. Improving statistical keyword detection in short texts: entropic and clustering approaches / C. Carretero-Campos, P. Bernaola-Galván, P. Ch. Ivanov, P. Carpena // Phys. Rev. E. – 2012. – Vol. 85. – 011139 (6 pp.).
38. Beliga S. Keyword extraction a review of methods and approaches / S. Beliga // Rijeka : Department of Informatics, University of Rijeka, 2014. – 9 p.
39. Onan A. Ensemble of keyword extraction methods and classifiers in text classification / A. Onan, S. Korukoğlu, H. Bulut // Expert Systems With Applications. – 2016. – Vol. 57. – P. 232–247.
40. Gupta E. T. Keyword extraction: a review / E. T. Gupta // Int. J. Eng. Appl. Sci. Technol. – 2017. – Vol. 2, No 4. – P. 215-220.
41. Kantelhardt J. W. Fractal and multifractal time series / J. W. Kantelhardt. – In: Mathematics of Complexity and Dynamical Systems. Ed. by R. A. Meyers. – New York : Springer, 2012. – P. 463–487.
42. Mosaic organization of DNA nucleotides / C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, A. L. Goldberger // Phys. Rev. E. – 1994. – Vol. 49, No 2. – P. 1685–1689.
43. Albert R. Statistical mechanics of complex networks / R. Albert, A.-L. Barabasi // Rev. Mod. Phys. – 2002. – Vol. 74, No 1. – P. 47–97.