Питання на іспит з навчальної дисципліни «*Опрацювання природної мови*»

1. Лінгвістика та її структура.
2. Загальні поняття про мову. Мова і мислення. Мова як знакова система. Мова і мовлення. Мовна структура. Рівні мови.
3. Фонологія. Поняття про фонеми. Звуки і букви. Поняття про алфавіт.
4. Морфологія. Основні поняття. Морфема. Види морфем.
5. Синтаксис. Основні поняття. Словосполучення. Речення. Синтаксичні відносини. Керування. Узгодження. Комунікативна організація речення.
6. Семантика. Семантичні відносини. Парадигматичні та синтагматичні відносини. Формальні моделі семантики. Семантичні мережі. Фрейми.
7. Лексикографія. Основи комп'ютерної лексикографії.
8. Прагматика. Основні поняття. Проблема розуміння текстів.
9. Історична (компаративна) лінгвістика. Синхронія і діахронія.
10. Основні родини мов. Приклади взаємних впливів.
11. Контрастивна лінгвістика або лінгвістична топологія.
12. Соціолінгвістика та діалектологія. Психолінгвістика.

13. Зв'язки комп'ютерної лінгвістики із системами штучного інтелекту та складними системами.
14. Основні завдання опрацювання природної мови.
15. Комп'ютерна лінгвістика. Квантитативна, статистична та математична лінгвістика.
16. Комп'ютерна лексикографія.
17. Приклади «lingware».
18. Контекстно-вільна граматика. Трансформаційна граматика.
19. Теорія «зміст–текст» у лінгвістиці.
20. Поняття «слово» в лінгвістиці. Словоформа, слововживання і лексема. Типи та токени.

21. Поняття систем. Системи із взаємодіючими елементами. Складні системи і мережі. Приклади. Лінгвістичні системи та мережі.
22. Статистичні підходи в лінгвістиці. Детермінізм і стохастичний підхід. Випадкові величини. Дискретні та неперервні величини.
23. Генеральна сукупність і вибірка. Репрезентативність вибірки. Ефекти «скінченного розміру» вибірки. Середнє значення та дисперсія вибірки.

24. Розподіли ймовірності випадкових величин. Гістограми. Масова функція розподілу ймовірності. Густина розподілу ймовірності.
25. Кумулятивна функція розподілу. Приклад однорідного розподілу.
26. Частотні таблиці слів. Алгоритмізація побудови частотних лексичних таблиць.
27. Статистичні моменти розподілів. Проблема існування моментів. Середнє значення випадкової змінної для «несправедливих» розподілів.
28. Середнє та середньоквадратичне відхилення для неперервних і дискретних випадкових величин. Зважені параметри.
29. Приклади розподілів випадкових величин. Зріст людини. Населення міст.
30. Приклади розподілів випадкових величин. Сила землетрусів. Розподіл багатства між людьми. Відвідування сайтів.
31. Частота слів і букв. Нульові гіпотези в статистичній лінгвістиці.
32. Розподіл Пуасона.
33. Експоненційний і «розширений» експоненційний розподіли.
34. Розподіл Вейбуля. «Логарифмічні» рангові залежності.
35. Нормальний розподіл. Причини універсальності нормального розподілу. Правило «трьох сигм».
36. Лог-нормальний розподіл. «Оманливі» степеневі розподіли.

37. Розподіл Ґуда. Частоти речень і слів різних довжин.
38. Розподіли з «важким хвостом». Степеневі розподіли. Розподіл Парето.
39. «Справедливі» та «несправедливі» розподіли. Правило Парето.
40. Степеневі розподіли. Скейлінг. Візуалізація степеневих розподілів.

41. Препроцесинг текстів. Парсинг.
42. Тегування, стеммінг і лемматизація.
43. Методика досліджень степеневих розподілів. Шуми функції $p(x)$. Бінування. Типи бінування.
44. Алгоритми визначення показника степеневих розподілів. Нелінійна апроксимація.
45. Метод максимальної схожості в оцінюванні показника степеневої залежності.
46. Критерій хі-квадрат у статистичній лінгвістиці. Інші статистичні критерії та тести.

47. Рангові залежності частоти слів. Абсолютна та відносна частота.
48. Перший і другий закони Ціпфа для лексики. Поняття про лексичний спектр.
49. Кумулятивна ймовірність появи слів з різними частотами. Закон Парето.
50. Лексичний словник. Закон Гіпса–Гердана.
51. Синтетичні та аналітичні мови. Порівняння словників синтетичних і аналітичних мов.
52. Метод біжучого вікна для розрахунку залежності словника від розмірів тексту.
53. Корпуси текстів. Закон Гіпса для корпусів текстів..

54. Альтернативи степеневих законів Ціпфа в лінгвістиці. Логарифмічні та експоненційні рангові розподіли і їхня графічна візуалізація. Ієрогліфічні мови з обмеженим словником.
55. Рангові залежності для частоти букв і типів складів. Залежність розмірів «словника» букв від довжини тексту. Рангові залежності в спортивних турнірах.
56. Взаємні зв’язки експонент степеневих законів для лінгвістики. Ілюстрації для статистики букв і східних мов.
57. Приклади статистичних розподілів поза межами лінгвістики. Розподіли ймовірності для імен і прізвищ. Прізвища в Кореї. Популярність наукових журналів, статей і авторів. Індекс Герша.

58. Показник степеневого розподілу лексичної ймовірності для текстів і проблема психічних захворювань. Особливості лексичних спектрів для дітей.
59. Основні теорії для пояснення законів Ціпфа. Нульові стохастичні моделі. «Фазовий перехід» комунікація–відсутність комунікації.
60. Закони лінгвістики для текстів програм.
61. Особливості статистики графем і фонем. Рангові та частотні залежності для букв і знаків. «Шуми» в рангових залежностях. «Словник» літер і закон Гіпса для літер.
62. n-грами в лінгвістиці. Буквені, символьні та лексичні n-грами. Узагальнення законів статистичної лінгвістики для n-грам.
63. Рангові та частотні залежності для слів різних довжин. Середня довжина слова в різних мовах. Особливості флуктуацій частоти літер.
64. Інформаційна ентропія Шеннона. Ентропії для залежностей $F(r)$ і $p(F)$. «Довжина тексту» з урахуванням ентропії.
65. Закони, що пов’язують лінгвістичні елементи різних рівнів. Закон Менцерата–Альтмана.
66. Розподіли ймовірності першої появи слів і лексичних n-грам у тексті.

67. Рандомні тексти. Типи рандомних текстів.
68. Тексти Хомського та Саймона. Модель зростання Саймона. Модифіковані моделі Саймона.
69. Тексти «мавпи Міллера». Тексти на основі ланцюжків Маркова.

70. Рандомізовані тексти. Локальна та глобальна рандомізація. Рандомізація та кореляції в текстах.
71. Властивості рандомних текстів. Проблема розпізнавання природних і рандомних текстів. Інвертування текстів.
72. Основні підходи в поясненні законів статистичної лінгвістики. Наближене відтворення законів Ціпфа, Парето та Гіпса.
73. Принцип «багатий стає багатшим» у складних системах.
74. Модель «дискурс–оточення». Поняття тригерування інновацій.
75. Принцип переважного приєднання в складних мережах.

76. Характеристика повторюваності в текстах. Основні означення.
77. Алгоритм та режими обчислень характеристики повторень.
78. Характеристика повторень для різних типів текстів. Пошук «самоплагіату» в текстах.

79. Порівняння і класифікація текстів. Лінгвістичні маркери.
80. Категоризація та класифікація текстів. Індексування текстів.
81. Проблеми визначення мови і встановлення авторства, жанру та стилю. Стилістика та її статистичні ознаки.
82. Проблема розпізнавання текстового плагіату. «Кусковий» плагіат.
83. Статистичні ваги маркерів текстів. Методика порівняння і класифікації текстів із машинним навчанням.
84. Параметри «recall» і «precision» в інформаційному пошуку.
85. Поняття подібності текстів і «відстаней» між текстами. Дефініції «відстаней». «Відстані» між ранговими залежностями і лексичними спектрами.
86. Опис подібності текстів на основі векторного простору. Скалярний добуток текстів.

87. Поняття ключових слів і фраз (n-грам). Частотні ознаки ключових слів.
88. «Абсолютні» та «відносні» методи визначення ключових слів.
89. Метод TF-IDF. Ключові слова в інформаційному пошуку. Метод Бріна та Пейджа.
90. Закон рідкісних подій і розподіл Пуасона.
91. Час очікування рідкісної події. Приклади. Аварії та технічні поламки, землетруси, серцеві ритми.
92. Час очікування рідкісної події в лінгвістиці. Експоненційний розподіл для часів очікування рідкісних подій.
93. Відхилення від експоненційного розподілу часів очікування рідкісних подій. Функціональні, змістові та ключові слова.
94. Семантика тексту. Розрізнення природних і рандомних текстів на основі кластеризації слів.
95. Ключові слова в текстах комп'ютерних програм.

96. Флуктуації. Приклади зі статистичної механіки. Відносні флуктуації.
97. Мікроскопічне, мезоскопічне і макроскопічне наближення в статистичній лінгвістиці.
98. Кореляції в статистичній лінгвістиці. Короткосяжні та довгосяжні кореляції.
99. Поняття кореляцій. Флуктуації за наявності кореляцій.
100. Поняття скейлінгу флуктуацій. Самоусереднювані статистичні величини.
101. Флуктуації в лінгвістиці. Відображення тексту на числовий ряд.
102. Закон Тейлора.
103. Часова мова в описі флуктуацій. Метод рандомних прогулянок. Нормальна та аномальна дифузія. Підхід біжучого вікна в дослідженні флуктуацій у лінгвістиці.
104. Ресурси флуктуаційного аналізу: кореляції, ключові слова та семантика тексту.
105. Флуктуації частоти слів у текстах програм.
106. Флуктуаційний аналіз на корпусах текстів.

107. Мережі. Основні параметри мереж.
108. Властивості складних мереж: перколяція, ефект тісного світу та механізм переважного приєднання.
109. Приклади складних мереж та їхні властивості.
110. Лінгвістичні мережі. Означення зв’язків між вузлами мережі.
111. Ефект «тісного світу» в лінгвістичних мережах.
112. Мережеві методи пошуку ключових слів текстів.