1. **Автоматический дискурсивный анализ (какие задачи входят, приведите примеры высокоуровневых задач, в которых необходим один из модулей автоматического анализа дискурса)**

Методология исследования дискурсивных феноменов, предложенная Пеше на основании моделирования механизмов социокультурной детерминации дискурсивных практик. Применяется в современных задачах автоматического извлечения информации из текста, такие как извлечение фактов, извлечение мнений, анализ контента на основе привлечения онтологических ресурсов.

2. Что такое расстояние Левенштейна между двумя строками? Каково расстояние

Это минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую.

Левенштейна между строками: мошка – мускат равно **3.**

3. **Что такое нормализация? Какие виды нормализации Вы знаете? Приведите пример, когда стемминг дает неудовлетворительные результаты.**
   Нормализация - это процесс трансформации текста в более каноничную форму, которая отличается от исходной.
   Виды нормализации:
   a. Лемматизация
   b. Стемминг

Неудовлетворительные результаты.
«universal», «university» и «universe» с основой «univers». Значения относятся к различным областям, поэтому рассматривать их как синонимы неверно

4. **Что такое лексическая вероятность в модели HMM для разрешения неоднозначности? Каково основное допущение относительно лексической вероятности в модели? Приведите пример, когда вклад лексической вероятности должен быть значимо высоким**

   Это установление конкретного значения слова в некотором контексте. Разрешение лексической многозначности является одной из центральных задач обработки текстов. Оно используется для повышения точности методов классификации и кластеризации текстов, увеличения качества машинного перевода, информационного поиска и других приложений.

   - **Основное допущение**: Вероятность увидеть некоторое слово в тексте зависит только от его собственного грамматического тега (от его собственной грамматической характеристики)

Вклад лексической вероятности должен быть значимо высоким когда частота слова как одной части речи, допустим, мала, а а другой велика!

Например: частота *saw* как существительного 4 раза на весь Брауновский корпус, а как глагола – 337 раз.

5. **Приведите примеры 2-х проблем для синтаксического анализа в терминах деревьев зависимостей**
    1. Дальние связи (согласование, модели управления)
    2. Нули (эллипсис, нулевые связки)

Часть 2.

6. **С помощью информации из НКРЯ рассчитайте, какая (условная) вероятность выше: вероятность «увидеть» существительное после наречия или вероятность “увидеть” существительное после прилагательного**

Общий объем 283 431 966
После Adv: 12 822
После Adj: 5 208

P(Adv) = 12 822 / 283 431 966
P(Adj) = 5 208 / 283 431 966

8.
John(N) built(V) a(Det) house(N).
John(NP) built(V) (a house(NP)).
(John built(VP)) (a house(NP)). S

He(N) ran(V) away(Part).
He(NP) (ran away(VP)). S

John(N) gave(V) Mary(N) roses(N).
John(N) gave(V) Mary(NP) roses(NP).
John(NP) (gave Mary roses(VP)). S

The(Det) curse(N) has(Aux) come(V) upon(Prep) me(Pron).

(The curse(NP)) has(Aux) come(V) upon(Prep) me(NP).

(The curse(NP)) has(Aux) come(V) (upon me(PP)).

(The curse(NP)) (has come upon me(VP)). S

Старался более менее понятно написать, надеюсь удобно читать.

КС-грамматика тогда будет иметь следующий вид:

S → NP VP
S → VP
NP → Det N
NP → N
NP → Pron
VP → V NP
VP → V Part
VP → V NP NP
VP → Aux V PP
PP → Prep NP

Det → a / the
N → John / house / Mary / roses / curse
Pron → me
V → built / ran /gave / come
Part → away
Aux → has
Prep → upon