**Ответы на тестовые задания с пояснениями**

**Вопрос 1.**

*Ранжирование — это ...*

**Правильный ответ:** Процесс задания частично упорядоченного множества

Вариант с "упорядоченным множеством" не подходит, потому как некоторые документы вообще могут не иметь отношений между собой, и невозможно сказать, какой из них находится выше или ниже при ранжировании (они вне шкалы).

Вариант "процесс определения схожести текстов документов к запросу" узок, так как в общем случае мы не говорим про тексты (можем ранжировать что угодно).

Вариант "определения границ схожести двух объектов" в целом не имеет смысла, понятие границ схожести не вводилось в лекции.

**Вопрос 2.**

*Дан некоторый объект A и набор объектов x=B,C,D,E. Объект B похож на A на 3 условных единицы. Объект C похож на A на 0.1 условную единицу. Объекты D и E не похожи на объект A. Можно ли назвать x упорядоченным множеством?*

**Правильный ответ**: Нет

Данный пример нельзя назвать упорядоченным множеством, так как существуют такие пары объектов, между которыми не установлены отношения, а значит множество упорядочено лишь частично.

**Вопрос 3.**

*Дан некоторый объект A и набор объектов x=B,C,D,E. Объект B похож на A на 3 условных единицы. Объект C похож на A на 0.1 условную единицу. Объекты D и E не похожи на объект A. Можно ли назвать x частично упорядоченным множеством?*

**Правильный ответ:** Да  
см. предыдущее объяснение.

**Вопрос 4.**

*С точки зрения обучения модели задача (Learning To Rank) является:*

**Правильный ответ:** Supervised задачей и Semi-supervised задачей. Обучение с учителем - классическое предсказание релевантности, взятое из некоторой **разметки.** Semi-supervisedслучай возможен, к примеру, если производилось дообучение на предсказаниях модели предыдущего уровня (обученной в Supervised-режиме).

Без меток модель ранжирования работать не может, так как не ясно, как вообще расставлять объекты. Любое введение отношений между объектами конвертируется в обучение с учителем (так как появляются явные целевые значения) - так что вариант Unsupervised некорректен. Multiderivative же вообще относится скорее к методам решения конкретной оптимизационной задачи.

**Вопрос 5.**

*Можно ли решать задачу ранжирования с точки зрения классификации?*

**Правильный ответ:** Да, задачу ранжирования можно решать как задачу классификации (релеванто/нерелевантно), однако, как было отмечено в лекции, это не лучший существующий способ.

**Вопрос 6.**

*Глобальная ранжирующая модель отличается от ранжирующей модели тем, что:*

**Правильный ответ:** Глобальная модель ранжирует сразу набор документов. Объяснение дано в лекции - отдельная модель обрабатывает конкретную пару запрос-документ, но над всем множеством документов строится глобальная модель, которая может состоять из разных частей (отсечение кандидатов и прочее)

Так что эти модели точно отличаются, и вариант "они не отличаются, это одно и то же" не подходит. Суждение "глобальная модель работает хуже, но быстрее" не верно, так как обработка всех документов как минимум не может быть быстрее одного документа. Вариант "Глобальная модель способна работать на новых данных" отсекается тем, что вообще любая хорошая модель (в том числе отдельной оценки запроса и документа) должна обладать некой обобщающей способностью, а значит и работать на новых данных.

**Вопрос 7.**

*Отметьте все варианты, в которых используется матчинг:*

**Правильный ответ**:

Поиск пары кроссовок в интернете (гугл/яндекс)

Нахождение товаров из вашего магазина “Трубы-и-сантехника” на Яндекс.Маркете

Нахождение одинаковых картинок в базе изображений (например, ImageNet)

Все это - верные и, кажется, очевидные варианты работы матчинга, когда несколько объектов сопоставляются друг с другом - будь то кроссовки, товары магазина или изображения (одного и того же стула, например, или автомобиля).

Вариант "Поиск всех вершин в графе с заданным свойством" не подходит потому, что здесь нет как такового процесса матчинга - происходит просто фильтрация по условию if/else.  "Поиск телефона для подруги на день рождения" опирается скорее на нечто внутреннее, вероятно, представление о том, каким должен быть подарок, но никаких явных алгоритмов машинного обучения здесь не запускается.

**Вопрос 8.**

*Отличие матчинга от ранжирования:*

**Правильны ответ:**

возможность ничего не выдавать в качестве вердикта (простой пример - когда мы не уверены, что вообще имеем похожие товары, и отдавать всякий мусор невыгодно)

в качестве выходного модуля в матчинге, как правило, стоит классификатор - это одно из основных отлчиий, разобранных в лекции. Это позволяет безотносительно абсолютных значений релеватности оценивать уверенность (вероятность) соответствия объектов, и не выдавать пары, в которых алгоритм не уверен.

Вариант "Матчинг применим к товарам, а ранжирование — к текстовым документам" не подходит, так как слишком узок (см. вопрос 1). Опция "Матчинг это ручная задача, а ранжирование — автоматизация того же процесса" не является корректной потому, что матчинг, как минимум, тоже можно автоматизировать.

**Вопрос 9.**

*Выберите корректное и полное определение матчинга:*

**Правильный ответ**: Матчинг - это процесс сопоставления объектов путем сравнения и расчета некоторой меры схожести. Можно пойти от обратного:

Матчинг это не процесс сбора данных (скорее процесс обогащения данных установлением связей).

Матчинг как инструмент сравнения цен - опять слишком узкое и неполное определение, равно как и "поиск текста документа, ближайшего по смыслу к запросу пользователя".

**Вопрос 10.**

*Мультимодальные модели это:*

**Правильный ответ:** Мультимодальные модели это модели, принимающие на вход данные разной природы и создающие внутри себя совместное скрытое представление (по определению).

Понятие мультимодальности не накладывает никаких ограничений на распределение входных объектов и выходных предсказаний, а значит варианты с Бимодальными распределениями некорректны.

Вариант "Модели, умеющие за раз обрабатывать несколько объектов (например, классифицировать целое предложение, состоящее из десятка слов)" никак не соотносится с понятием природы данных.

**Вопрос 11.**

*Примерами мультимодальных моделей являются:*

**Правильный ответ:**

система ответа на вопрос по картинке (потому что нужно и понять вопрос, то есть текст, и обработать картинку, то есть визуальное представление)

система оценки токсичности мемов и шуток в комментариях (мемы и шутки тоже бывают в виде и текстов, и картинок)

Не подходят варианты:

встроенный поиск Википедии - если вы им пользовались, то знаете, что он ищет по тексту (одна природа данных).

система распознавания лиц для отлова преступников - работает только с изображениями

языковая модель для предсказания оценки по отзыву на IMDB - обрабатывает только текст отзыва

**Вопрос 12.**

*Выберите проблемы, присущие задаче матчинга:*

**Правильный ответ:** все указанное, кроме "ничего из указанного" :)

"опечатки" - человек неправильно ввел название товара, и его сложнее сматчить (сопоставить "карандащ" и "карандаш")

"сокращения в названиях"  и "отсутствие стандарта формирования названия товара" - примеры приведены в лекции.

"Большая схожесть между разными товарами" - основная проблема матчинга (общая схожесть товаров и отличие в одном-двух символах)

**Вопрос 13.**

*Выберите, что подходит под определение SKU:*

**Правильный ответ:** 34759205759 , iPhone 7S 64GB красный (FG283X) - позволяет явно выделить конкретный товар по уникальному коду / артикулу.

"iPhone 8S (2017)" в противовес является обобщением, как минимум в поколении смартфонов были разные цвета. "Бананы 1 кг" - могут быть разные бананы (разные размеры, вы видели маленькие бананы?)

**Вопрос 14.**

*Выберите, какие пары запросов-документов указаны верно:*

**Правильный ответ**: все варианты корректны, "всё вышеперечисленное"

Гугл и страницы из интернета - это классическое поисковое ранжирование, с присущим соответствием запросов и документов.

Ассортименты магазина - это матчинг одной базы товаров на другую. Не важно, какая из них запросы, а какая - документы. Как и сопоставление моделей автомобилей с каталогом это - сопоставление некоторой сущности с базой таких сущностей.

**Вопрос 15.**

*Укажите, каким может быть тип связи модели и SKU в реальной жизни:*

**Правильный ответ**: Все вышеперечисленное.

Основные связи в идеальном мире были рассмотрены в лекции. А связь "Многие модели к одному SKU" запросто может встретиться в реальной жизни, если ваши данные не дедуплицированы (люди, заводящие товары, могут просто не уследить, и завести схожий товар, являющийся полным клоном). Будьте аккуратны и внимательны, когда делаете join-операции между таблицами-справочниками - можете словить неприятное размножение данных.

**Вопрос 16.**

*Какие эвристики можно применить для отсечения больших частей базы документов при решении задачи матчинга товаров на маркетплейсе?*

**Правильный ответ:**

Выделять производителя товара и работать только в рамках него - сужаем круг документов матчинга буквально за бесплатно, без запуска тяжелейших расчетов. Обычно можно выделить производителя регулярным выражением или заранее заготовленным словарем (с поиском подстроки)

Проверять, что запрос является товаром, а не запросом в обычный поисковик - чтобы не тратить электричество почём зря, если спрашивают даже не про товар (так, например, делает Яндекс)

Определять категорию товара и работать только с ней - аналогично первому пункту

Вариант "Случайным образом выбрать по 20 товаров каждого цвета и ранжировать только их" глуп и не имеет смысла - мы легко можем не выбрать среди двадцатки нужный товар, и упустить его в выдаче.

Вариант "Запускать глобальную ранжирующую модель на всех данных, а затем часть запросов подавать на вход ранжирующей модели" описывает полный пайплайн с запуском расчётов, и не может называться простой эвристикой и вызывать ускорение.