Анализ неструктурированных данных
Самостоятельная работа
Максимальная оценка: 10 баллов

Имя и фамилия: _____

Группа: _____

Ответьте на вопросы ниже. Можно использовать компьютер для вычислений и для просмотра лекций и учебников. Запишите решения, выделите ответы и прокомментируйте каждый шаг решения.

Каждый вопрос оценивается в 2 балла.

1. Пусть дан следующий фрагмент статьи из Википедии: "*Пик предпринимательской деятельности Саввы Мамонтова пришёлся на последнее десятилетие XIX века, когда он начал осуществлять Северный железнодорожный проект.*" Что из перечисленного ниже является решением задачи NER?

   A [Саввы Мамонтова = он]

   B [Саввы Мамонтова, Северный железнодорожный проект]

   C [пик, деятельность, савва, мамонтов, прийтись, последний, десятилетие, век, осуществлять, северный, железнодорожный, проект]

   D [Савва Иванович Мамонтов = он, осуществлять, Северный железнодорожный проект]

   Ответ: B

2. Рассмотрим модель CNN-biLSTM-CRF. Ее первая часть, char-CNN, помогает:

   A учесть регистр и особенности словообразования

   B учесть порядок слов

   C обработать слова не из словаря (OOV)

   D разрешить омонимию и полесимию

   Ответ: A, C

3. Продолжим рассматривать модель CNN-biLSTM-CRF. Ее вторая часть, word-BiLSTM,

   A не может быть использована как самостоятельный алгоритм sequence labelling

   B нужна для вычисления распределенных представлений слов

   C требует экспертного составления признакового пространства

   D использует одновременно два вида нейронов: LSTM и GRU нейроны

   Ответ: B

4. Продолжим рассматривать модель CNN-biLSTM-CRF. Ее третья часть, CRF,

   A не может быть использована как самостоятельный алгоритм sequence labelling

   B нужна для вычисления распределенных представлений слов

   C нужна для глобального перевзвешивания выхода BiLSTM

   D в среднем имеет бОльшее, по сравнению с BiLSTM, количество параметров

   Ответ: C

5. Существует два базовых варианта оценки качества NER: по токенам и по сущностям. Какой из этих вариантов дает большее значения accuracy?

   Ответ: по токенам