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.**

1. **Языковая модель [6 баллов]**

   Рассмотрим маленькую коллекцию документов, состоящую из следующих документов:

   1. <s> I am Sam </s>
   2. <s> Sam I am </s>
   3. <s> Sam I like </s>
   4. <s> Sam I do like </s>
   5. <s> do I like Sam </s>

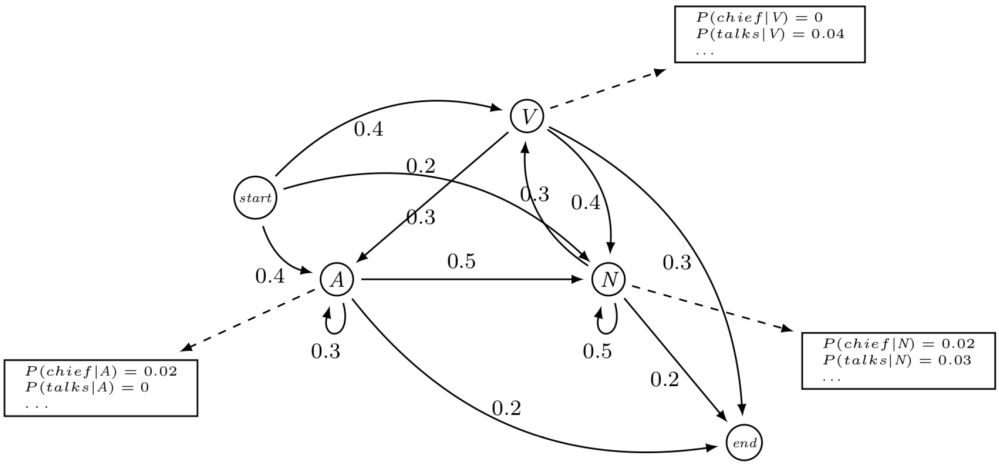
   По этой коллекции обучается модель биграм.

   (a) Чему равна $P(</s>|\text{like})$?
   ОТВЕТ: $P(\text{like } </s>)$  $P(\text{like}) = 2\ /3$

   (b) Какое слово наиболее вероятно после последовательности <s> Sam ... ?
   ОТВЕТ: I

   (c) Какой документ будет найден по запросу <s> I like </s> ?
   ОТВЕТ: 3) (самый короткий, из содержащих запрос)

2. **Скрытые цепи Маркова [4 балла]**




   (a) Чему равна вероятность последовательности слов с тегами N V ?
   ОТВЕТ: $P$(start, N) * $P$(N, V) * $P$(V, end) = 0.2 * 0.3 * 0.3 = 0.018

   (b) Чему равна P(chief talks, N N)?
   ОТВЕТ: $P$(start, N) * $P$(chief, N) * $P$(N, N) * $P$(talks, N) * $P$(N, end) = 0.2 * 0.02 * 0.5 * 0.03 * 0.2 = 0.000012