***Комментарии к заданию по Topic Modeling.***

1. Для задания был выбран датасет на английском языке 4\_en. Он оказался небольшим, 4276 документа, медиана длины в символах – 164. Остальные настройки остались такими же, как и в примере на лекции (20 топиков, униграммы). Тематическое моделирование осуществлялось с помощью методов LDA (sci-kit), NMF (sci-kit) и LDA (gensim). Результаты сразу дали понять о том, что представленные данные – это новостные статьи, причем американские. Это стало понятно после того, как в одном из топиков по NMF в топе слов встретились “great, make, america”, а в других – “trump” и “military, isis”. Т.е. статьи по политической тематике, в основном посвященные последним выборам в США и ИГИЛ. Выдача LDA (sci-kit), в целом, похожа на выдачу NMF, однако, по моему мнению, NMF сработал более качественно, темы статей угадываются легче. Однако метод LDA от gensim (поданные на вход тексты дополнительно были очищены от nltk-стоп-слов) справился с задачей, по моему мнению, лучше двух других методов – темы были более эксплицитны и угадывались чуть легче. Так как датасет был небольшим, gensim-LDA работал буквально пару секунд.
2. На втором этапе было решено попробовать использовать биграммы для LDA (sci-kit) и NMF (sci-kit) и сравнить результаты с «победившим» на предыдущем этапе LDA (gensim). Все остальные параметры остались без изменений. В итоге, модели с использованием биграмм оказались намного эффективнее модели с униграммами, кроме того, небольшая разница в качестве между LDA (sci-kit) и NMF (sci-kit) нивелировалась, тематики легко угадывались в обоих случаях.

В общем и целом, все методы показали достойные результаты, обработка не заняла много времени, но стоит также отметить и то, что датасет был небольшим, что не позволило оценить разницу между методами в скорости работы. На униграммах лучшие результаты показал LDA от genism, на биграммах LDA (sci-kit) и NMF (sci-kit) оказались одинаково хороши.