Выводы по лабораторной работе №3

В работе были протестированы три стратегии выбора обучающих данных:

1. Случайная выборка
2. Алгоритм активного обучения Least Confidence
3. Алгоритм активного обучения Margin Sampling

Результаты: *results.csv* и *results.png*

1. 100% обучающего датасета

Полное обучение на 100% обучающего датасета дало качество:

F1-macro = 0.9134

2. Случайная выборка: 1%, 10% и 20% обучающего датасета

Доля данных	Средний F1-macro
1%	0.8500
10%	0.8960
20%	0.8987

Можно заметить, что уже 10% данных дают качество, почти совпадающее с обучением на 100%. Увеличение с 10% до 20% даёт минимальный прирост.

3. Алгоритм активного обучения Least Confidence

Доля данных	Итоговый F1-macro
1%	0.7362
10%	0.8244
20%	0.8495

LC показал результаты хуже случайной выборки на всех уровнях. Наиболее заметное отставание при 1% данных: 0.736 против 0.85 у случайной выборки.
На мой взгляд, на это есть следующие причины:

- Модель на старте имеет мало данных и плохо оценивает уверенность.

- Алгоритм выбирает слишком сложные примеры, которые не дают хорошего обобщения.

- Для чистых и сбалансированных датасетов (каким и является AG News) случайный выбор оказывается более эффективным.

4. Алгоритм активного обучения Margin Sampling

Доля данных	Итоговый F1-macro
1%	0.8351
10%	0.8965
20%	0.8994

При 1% данных Margin Sampling уже почти догоняет случайную выборку: 0.835 против 0.85. На 10% данных метод даёт 0.8965, что практически совпадает со случайной выборкой. На 20% данных Margin Sampling показывает лучший результат среди всех методов кроме полного обучения: 0.8994.

Общий вывод

Метод Margin Sampling является наиболее эффективным из протестированных методов активного обучения: он позволяет достичь качества, сравнимого со случайной выборкой, но потенциально при значительно меньших затратах на разметку.

Метод Least Confidence на данном датасете ухудшает качество, что подтверждают результаты эксперимента и визуализация.