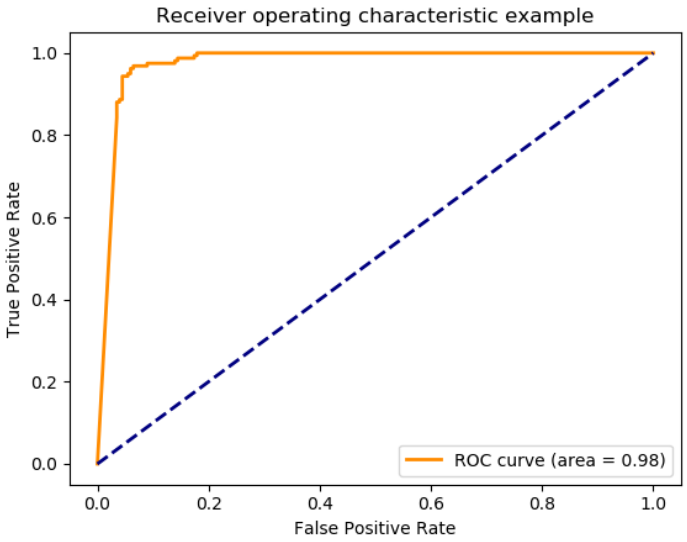
Классификация текста

Набор данных

- Набор данных содержит множество сообщений (файлов).
- Если в названии файла есть подстрока **legit**, то это **хорошее** сообщение.
- Если в названии файла есть подстрока **spmsg**, то это **спам**.
- Сообщения состоят из заголовка и тела сообщения.
- Каждое слово было заменено на определённое число.

Задание

1. Придумайте два способа учёта заголовка и тела сообщения при векторизации. Например: учитывать одинаково или учитывать раздельно.
2. Придумайте два способа векторизации текста. Например, для каждого слова создать признак: встретилось или нет, или сколько раз встретилось.
3. Табличный набор данных должен быть в разреженном виде.
4. Постройте Наивный байесовский классификатор на наборе данных.
5. Постройте ROC кривую. Вычислите для неё AUC. Выберете наилучшую комбинацию из первых двух пунктов относительно метрики AUC.



6. Измените априорное распределение для выбранной модели таким образом, чтобы число **хороших** сообщений классифицированных как **спам** равнялось нулю.
7. Постройте график зависимости точности классификации от изменения априорного распределения в логарифмированном пространстве.