Практическое занятие № 7. Вернемся к Титанику

На этом занятии мы возвращаемся к набору данных о выживших пассажирах на Титанике.Этот набор данных был опубликован по ссылке http://bit.ly/2ciHWiw. Он записан в формате csv и кроме всего прочего содержит вот такие атрибуты:

- **PassengerId**, уникальный в пределах файла идентификатор пассажира;
- **Survived**, 0 если пассажир не выжил и 1 если выжил после крушения;
- **Pclass**, класс (от 1 до 3), которым путешествовал пассажир;
- **Last Name**, фамилия пассажира;
- **First Name**, имя пассажира;
- **Sex**, пол пассажира (F — женский, M — мужской);
- **Age**, возраст пассажира;
- **Fare**, стоимость билета.

С другой стороны, вот эта статья в Wikipedia содержит еще один список пассажиров Титаника: https://en.wikipedia.org/wiki/Passengers_of_the_RMS_Titanic.

1. Задачи

Перед вами стоит проблема понимания, существуют ли расхождения в двух наборах данных и если да, то какие это расхождения. Для этого постарайтесь выполнить следующие задачи:

1. напишите код, который читает csv-файл и помещает его в список (используйте модуль csv);
2. напишите код, который автоматически скачивает страницу из Wikipedia (используйте модуль requests);
3. напишите код, позволяющий распарсить скачанную html-страницу вытаскивает из нее необходимые данные (модуль lxml + вспомните XPath);
4. выявите расхождения между источниками: какие персоны есть в одном источнике, но их нет в другом;
5. попробуйте использовать *нечеткое* сравнение между именами персон.

2. Нечеткое сравнение

Пусть даны две строки: `Strandberg, Miss. Ida Sofia` и `Strandberg, Miss Ida Sofia`. Они отличаются только тем, что внутри одной строки есть точка, а в другой строке точки нет.

Разобьем эти строки на множества подстрок длины два: A ={'St', 'tr', 'ra',...,'Mi', 'is', 'ss', 's.', '. ',...} и B={'St', 'tr', 'ra',...,'Mi', 'is', 'ss', 's ',...}.

Чтобы понять, являются ли два множества похожими можно посчитать меру Жаккара:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Чем выше значение $J(A,B)$, тем больше общих элементов есть в подмножествах.

Поможет ли использование этой меры для решения нашей задачи?