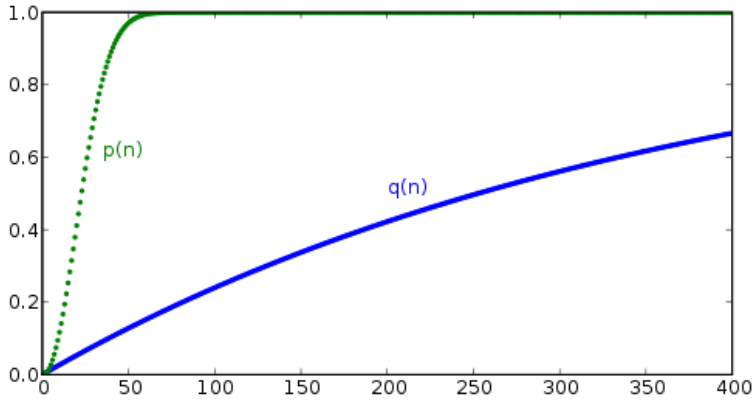
Birthday paradox

It's puzzling but true that in any group of 23 people there is a 50% chance that at least two of them share the same birthday. This **birthday paradox** isn't a logical paradox — there's nothing self-contradictory about it — it's just unexpected.



The birthday paradox: $p(n)$ indicates the probability that in a group of n people at least two of them share the same birthday; $q(n)$ indicates the probability that in a group of n people there is at least one person sharing the same birthday as another person that was chosen beforehand (for example: yourself). The value n is shown on the horizontal axis and probabilities (in the interval from 0 to 1, or the interval from 0% to 100%) are shown on the vertical axis.

Imagine a typical classroom having 30 students. People usually think it's an amazing coincidence that two students in such a class share the same birthday, but is isn't all that rare after all. Actually, with 30 students there is a 70% chance.

Perhaps the best data set of all to test the birthday paradox could be found this summer on the 2014 FIFA World Cup, an international football tournament held in Brazil from 12 June to 13 July 2014. The 32 national teams involved in the tournament were required to register a squad of 23 players, including three goalkeepers. Only players in these squads were eligible to take part in the tournament. If the birthday paradox is true, 50% of the squads should have shared birthdays.

In order to test this, we have collected information about all players from FIFA's official squad list. We have stored the information about all players from a national team squad in a text file. Each line of in such a file contains the following information fields for a single player: *i*) name, *ii*) national team, *iii*) squad number, *iv*) position on the field (GK=goalkeeper, DF=defender, MF=midfielder, FW=forward), *v*) date of birth (YYYY-MM-DD), *vi*) number of caps and *vii*) club. Information fields are separated using commas (,). As an example, the first few lines of the file france.txt that contains information about the players of the French national team are shown below.

```
Hugo Lloris,France,1,GK,1986-12-26,57,Tottenham Hotspur
Mathieu Debuchy,France,2,DF,1985-07-28,21,Newcastle United
Patrice Evra,France,3,DF,1981-05-15,58,Manchester United
Raphaël Varane,France,4,DF,1993-04-25,6,Real Madrid
Mamadou Sakho,France,5,DF,1990-02-13,19,Liverpool
Yohan Cabaye,France,6,MF,1986-01-14,30,Paris Saint-Germain
```

Assignment

In this exercise we process text files that contain information about all players in a national team squad on the World Cup football. The information about the players is stored in the format as outlined in the introduction. All text files use UTF-8 character encoding (see below). Your task:

- Write a function birthdays that takes the location of a text file. The function must return a dictionary that maps all **days in the year** in which at least one player in the national team squad celebrates its birthday onto the set of players who are born on that day of the year. Note that the keys of the dictionary are days in the year (format MM-DD) that need to be derived from the birthdays of the players as stored in the given file.
- Use the function birthdays to write a function birthdayparadox that takes the location of a text file. The function must return a Boolean value that indicates whether or not there is a day in the year in which at least two players of the national team squad celebrate their birthday.
- Use the function birthdayparadox to write a function testparadox that takes a list of national team squads. Each national team squad is represented as a tuple containing two strings: the name of the country and the location of the text files containing information about all players in the squad. The function must return a set containing the names of all countries whose national team squad has at least two players that celebrate their birthday on the same day of the year.

Unicode files

Unicode is a computing industry standard for the consistent encoding, representation and handling of text expressed in most of the world's writing systems. The standard consists of a set of code charts for visual reference, an encoding method and set of standard character encodings, a set of reference data computer files, and a number of related items, such as character properties, rules for normalization, decomposition, collation, rendering, and bidirectional display order (for the correct display of text containing both right-to-left scripts, such as Arabic and Hebrew, and left-to-right scripts).

Unicode can be implemented by different character encodings. The most commonly used encodings are UTF-8, UTF-16 and the now-obsolete UCS-2. UTF-8 uses one byte for any ASCII character, all of which have the same code values in both UTF-8 and ASCII encoding, and up to four bytes for other characters. To open a Unicode file in Python, you can pass the encoding used to the optional parameter encoding of the built-in function open. For example, in order to read the information from the Unicode text file france.txt that makes use of UTF-8 encoding, the file can be opened in the following way:

```
>>> open('france.txt', 'r', encoding='utf-8')
```

Example

In the following interactive session we assume that the text files algeria.txt, belgium.txt and france.txt are located in the current directory.

```
>>> born = birthdays('france.txt')
>>> born['02-13']
{'Mamadou Sakho', 'Eliaquim Mangala'}
>>> born['03-08']
{'Rio Mavuba', 'Rémy Cabella'}

>>> birthdayparadox('france.txt')
```

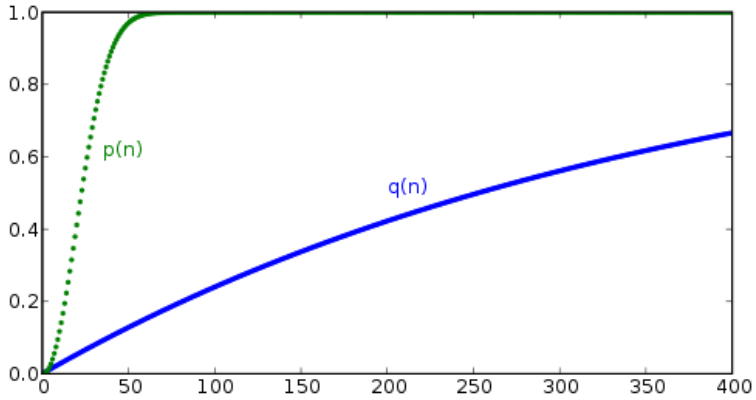
```
True
>>> birthdayparadox('belgium.txt')
False

>>> testparadox([('Algeria', 'algeria.txt'), ('Belgium', 'belgium.txt'), ('France', 'france.txt')])
{'Algeria', 'France'}
```

References

J. Fletcher (2014). The birtday paradox at the World Cup. BBC News Magazine.

De **verjaardagsparadox** stelt dat er meer dan 50% kans is dat in een groep van 23 personen, minstens twee personen dezelfde verjaardag hebben. Het gaat hierbij niet om een logische paradox — er valt nergens een tegenstrijdigheid te bespeuren — maar de observatie op zichzelf is contra-intuïtief.



De verjaardagenparadox: $p(n)$ geeft de waarschijnlijkheid dat in een groep van n personen er twee of meer op dezelfde datum jarig zijn; $q(n)$ geeft de waarschijnlijkheid dat in een groep van n personen ten minste één op dezelfde datum jarig is als een van tevoren gekozen persoon (bv. uzelf). Op de horizontale as n, op de verticale as de waarschijnlijkheid (van 0 tot 1, oftewel van 0% tot 100%).

Denk bijvoorbeeld aan een klas van 30 studenten. De meeste mensen denken dat het eerder toeval is als twee studenten van die klas op dezelfde dag jarig zijn. Maar dat is het allerminst. De kans dat het gebeurt is zelfs meer dan 70%.

De beste dataset om de verjaardagsparadox te testen, was deze zomer wellicht te vinden op de wereldbeker voetbal in Brazilië. Aan de eindronde van dit toernooi namen 32 landenploegen deel, die elk bestonden uit 23 spelers. Als de verjaardagsparadox waar is, dan verwachten we dat ongeveer 50% van de landenploegen minstens twee spelers hebben die op dezelfde dag jarig zijn.

Om dat te testen, hebben we uit de officiële ploegenlijst van de FIFA enkele gegevens over de spelers verzameld. We hebben daarbij de gegevens van alle spelers van een ploeg opgeslagen in een tekstbestand. Elke regel van zo een bestand bevat de volgende gegevens van een speler: *i*) naam, *ii*) landenploeg, *iii*) rugnummer, *iv*) positie op het veld (GK=doelman, DF=verdediger, MF=middenvelder, FW=aanvaller), *v*) geboortedatum (JJJJ-MM-DD), *vi*) aantal caps en *vii*) ploeg. Deze informatievelden worden telkens van elkaar gescheiden door een komma (,). Bij wijze van voorbeeld staan hieronder de eerste regels van het bestand france.txt dat informatie bevat over de spelers van het Franse elftal.

Hugo Lloris,France,1,GK,1986-12-26,57,Tottenham Hotspur
Mathieu Debuchy,France,2,DF,1985-07-28,21,Newcastle United
Patrice Evra,France,3,DF,1981-05-15,58,Manchester United
Raphaël Varane,France,4,DF,1993-04-25,6,Real Madrid
Mamadou Sakho,France,5,DF,1990-02-13,19,Liverpool
Yohan Cabaye,France,6,MF,1986-01-14,30,Paris Saint-Germain

Opgave

In deze opgave werken we steeds met tekstbestanden die informatie bevatten over alle spelers van een voetbalelftal op het wereldkampioenschap voetbal. Deze informatie over de spelers is opgemaakt volgens het formaat dat werd beschreven in de inleiding. Alle tekstbestanden gebruiken UTF-8 codering (zie hieronder). Gevraagd wordt:

- Schrijf een functie verjaardagen waaraan de locatie van een tekstbestand moet doorgegeven worden. De functie moet een dictionary teruggeven, die alle **dagen in het jaar** waarop minstens één speler van het elftal jarig is, afbeeldt op de verzameling van spelers die op die dag jarig zijn. Merk dus op dat de sleutels gevormd worden door dagen in het jaar (in het formaat MM-DD) die afgeleid moeten worden van de **geboortedatums** van de spelers uit het gegeven bestand.
- Gebruik de functie verjaardagen om een functie verjaardagsparadox te schrijven. Aan deze functie moet de locatie van een tekstbestand doorgegeven worden. De functie moet een Booleaanse waarde teruggeven die aangeeft of er een dag is waarop minstens twee spelers van het elftal jarig zijn.
- Gebruik de functie verjaardagsparadox om een functie testparadox te schrijven. Aan deze functie moet een lijst van landenploegen doorgegeven worden. Elke landenploeg wordt hierbij voorgesteld door een tuple met twee strings: de naam van het land en de locatie van een tekstbestand dat informatie over de spelers van de landenploeg bevat. De functie moet een verzameling teruggeven met de namen van alle landen die minstens twee spelers hebben die op dezelfde dag jarig zijn.

Unicode bestanden

Unicode is een internationale standaard voor de codering van binaire codes naar grafische tekens en symbolen, vergelijkbaar met de ASCII-standaard. De Unicode tekenset is echter veel uitgebreider dan de 256 symbolen van de ASCII-tekenset. Unicode ondersteunt een aantal mogelijke coderingen voor de tekenset, die aangeven hoe de symbolen binair voorgesteld worden. Als je in Python een Unicode bestand wilt openen, dan kan de gebruikte codering doorgegeven worden aan de parameter encoding van de ingebouwde functie open. Om bijvoorbeeld informatie te lezen uit het Unicode bestand france.txt dat gebruikt maakt van UTF-8 codering, kan het bestand als volgt geopend worden:

```
>>> open('france.txt', 'r', encoding='utf-8')
```

Voorbeeld

Bij onderstaande voorbeeldsessie gaan we ervan uit dat de tekstbestanden algeria.txt, belgium.txt en france.txt zich in de huidige directory bevinden.

```
>>> jarigen = verjaardagen('france.txt')
>>> jarigen['02-13']
```

```
{'Mamadou Sakho', 'Eliaquim Mangala'}
>>> jarigen['03-08']
{'Rio Mavuba', 'Rémy Cabella'}

>>> verjaardagsparadox('france.txt')
True
>>> verjaardagsparadox('belgium.txt')
False

>>> testparadox([('Algerije', 'algeria.txt'), ('België', 'belgium.txt'), ('Frankrijk', 'france.txt')])
{'Algerije', 'Frankrijk'}
```

Bronnen

J. Fletcher (2014). The birtday paradox at the World Cup. BBC News Magazine.