Patent Analytics Training Datasets

In this article we introduce the patent datasets developed for the Open Source Patent Analytics Project as training sets for patent analytics. The datasets will be used in the walkthroughs. The datasets will grow over time but we will briefly introduce them and explain how to access them.

The datasets are housed at the project [GitHub repository](https://github.com/poldham/opensource-patent-analytics/tree/master/2_datasets). The easiest way to access the files is is to download the complete set as a [zip file](https://github.com/poldham/opensource-patent-analytics/blob/master/2_datasets/datasets.zip?raw=true). To download individual files click on the link and then select raw to download the file.

## The datasets

The datasets are intended to illustrate the range of possibilities for the presentation of patent data including some of the challenges that may be encountered in cleaning and analysing patent data. They are also drawn from different sources.

### Pizza patent datasets

Almost everyone likes pizza and it is easy to search a patent database for the term "pizza". It is also an area of patent activity that encompasses a wide range of technologies such as pizza ovens, pizza boxes, pizza cutters and pizza toppings etc. It is therefore useful for demonstrating ways of interrogating patent data for particular topics.

1. pizza\_small is a very small dataset created by downloading the first page of results from the [European Patent Office espacenet database](http://worldwide.espacenet.com/?locale=en_EP) for a smart search on "pizza". It has only 26 rows. It's a quick and easy test dataset.
2. pizza\_medium was created from a search of the [WIPO Patentscope database](http://www.wipo.int/patentscope/en/) for the term "pizza" and contains 10,000 rows. It is intended to illustrate the data format from Patentscope and to allow work on a medium sized dataset. Note that the format varies from the espacenet format and presents different challenges.
3. pizza\_sliced is a set of five .csv files for a search of pizza on [espacenet](http://worldwide.espacenet.com/searchResults?ST=singleline&locale=en_EP&submitted=true&DB=worldwide.espacenet.com&query=pizza&Submit=Search). It is designed to illustrate issues involved in loading multiple files into R. It also illustrates problems with character corruption and the importance of pre-cleaning data before analysis.

### Patent Landscape Report datasets

Three datasets are drawn from the [WIPO Patent Landscape Reports](http://www.wipo.int/patentscope/en/programs/patent_landscapes/). The datasets address different topics, present a variety of fields and formats and are different sizes. Each dataset is linked to a detailed patent landscape report that provides an insight into approaches to patent analytics.

1. ewaste presents the results of research for a [report](http://www.wipo.int/patentscope/en/programs/patent_landscapes/reports/ewaste.html) on patent activity for electronic waste recycling and its implications for developing countries.
2. solar\_cooking presents the data supporting a [report](http://www.wipo.int/patentscope/en/programs/patent_landscapes/reports/solar_cooking.html) on technologies that use solar energy as the source for cooking and pasteurizing food.
3. ritonavir presents the data for a patent [report](http://www.wipo.int/patentscope/en/programs/patent_landscapes/reports/ritonavir.html) on patent activity for the HIV antiretroviral drug Ritonavir in the field of pharmaceuticals. The dataset illustrates specific activity around issues such as dosage and also the problem of 'evergreening' in patent activity.

### Other datasets

1. wipo is a single Excel sheet of data on trends in patent applications and growth rates from the [WIPO World Intellectual Property Indicators - 2014 Edition](http://www.wipo.int/ipstats/en/wipi/). The data is used for simple graphing in tools such as R and illustrates the need to skip rows when reading data into analytics tools.
2. WIPO\_sequence\_data. This dataset contains a small sample of the sequence data from the year 2000 available free of charge from the [WIPO Patentscope database](https://patentscope.wipo.int/search/en/sequences.jsf). This dataset can be used to explore analysis of patent sequence data.
3. [Synthetic biology](https://github.com/poldham/opensource-patent-analytics/tree/master/2_datasets/synbio_patents). This is a sample of data from Thomson Innovation developed by Paul Oldham for research on patent activity involving synthetic biology. The data has been extensively cleaned in VantagePoint from Search Technology Inc. and is intended to illustrate the use of data from a commercial patent database.

### Round Up

The datasets section of the project provides a series of useful training sets from a variety of sources and displaying a variety of features. These are open access datasets that can be used to test different approaches but please credit their sources. We will add more datasets over time and please feel welcome to submit datasets on specific topics of interest provided that they are not confidential. We are particularly interested in sample data from STN and QuestelOrbit or other data providers that can be used as training sets.