Identifying emerging trends from patent data

2018-04-12

Patents and other technical literature have key terminology trends identified, which may inform business and government decisions regarding new technologies. The analysis includes when and where terminology usage occurs, considered both nationally and internationally.

Team members

- Thanasis Anthopoulos
- Ian Grimstead
- Bernard Peat
- Emily Tew
- Michael Hodge

The need

Allow Cabinet Office (CO) and Department for Business, Energy & Industrial Strategy (BEIS) to identify popular technical terminology from patent abstracts to help their analysis of the innovation landscape in the UK and globally.

IPO is also interested in the above outcomes, but they are also interested in term emergence and in particular identifying emerging terminology as early as possible.

Finally, research teams within ONS are interested in forecasting emerging technical terms in order to be able to obtain and allocate resources to address future trends more efficiently.

Impact

The eventual output will be a number of apps that can address the stakeholders needs. In particular BEIS data-scientists have installed the technical term extraction module and the D3 visualisation into their servers and received

training from us on how to use these solutions with the view to make it available to their policy team.

The IPO has shown great interest in this project and we are working closely in collaboration with the data-science and analytics team in order to make the term-emergence module operational.

We also endeavour to make use of this module in ONS in order to early detect where state of the art statistical research is heading from research papers and plan Human Resources and dataset acquisition accordingly.

Data science

Machine Learning, Natural Language Processing (NLP), D3 for visualisation, cross-filter for multi-dimensional display.

Stakeholders

BEIS, CO, IPO, ONS, Python users who might wants to pick keywords out of patents.

Code and outputs

Data engineering:

- patent_reader - A utility module to create serialized objects from XML based source data into a consolidated python object. XML can be retrieved from local storage or bulk downloaded from the internet.
- app_bridge - A utility module to process patent data from sql tables into a into a consolidated python object.
- patloc.py: is the module that collates all the data needed to visualise where and when innovation takes place. It creates a csv file, which is consumed by our visualisation code written in js.

Problem Solving

- pat2gc.py is the module able to classify documents such as patents, papers and grants into subsets of interest.
- patent_app_detect - Open Github repository containing Python code for deriving popular terminology included within a particular patent technology area (CPC classification), based on text analysis of patent abstract information.

- Javascript D3 visual using the chiasm framework for the layout of different components, leaflet.js for the map component and chrossfilter.js for efficient display and filtering of multidimensional data.
- emtech.py is a module that can assess over a 10 year period, whether a term is classified as emergent or not and give an emergence score(ie. 'mobile phone' 101.332, '3d image' 33.45). This is based on a recent publication (Porter et.al 2018).
- emtech_predict.py (in progress) A module to forecast document counts for a 10 year period given the first few years as ground truth. For example, if self 'driving cars' have a 4-year history predict the next 6. The output can be processed by emtech.py to give an emergence prediction.

Delivery

- [x] **July 2018** Deliver popular terminology for a specified specified technology area
- [x] **September 2018** ML tool for classifying patent docs to a specified technology area
- [x] **October 2018** Visual to show where Innovation takes place
- [x] **October 2018** Identify Emerging Technologies
- [ ] **December 2018** Forecast Emerging Technologies

Further information

Please contact datasciencecampus@ons.gov.uk for more information.

Updates

- No updates yet.