Documentation USPTO patent text

This document describes data collection, processing, and different open access data files related to the text of USPTO patent documents. If you use the code or data, please cite the following paper:

Arts S, Hou J, Gomez JC. (2020). Natural language processing to identify the creation and impact of new technologies in patent text: code, data, and new measures. Forthcoming *Research Policy*. (https://doi.org/10.1016/j.respol.2020.104144)

Data and code

Data: https://zenodo.org/record/3515985 (DOI: 10.5281/zenodo.3515985)
Code: https://github.com/sam-arts/respol_patents_code
Paper: https://doi.org/10.1016/j.respol.2020.104144

Data processing

We collect patent titles, abstracts and claims for all granted U.S. utility patents from patentsview.org, the patent claims research dataset (Marco et al., 2016), and PATSTAT. We include all patents granted between March 1969 and May 2018 (n=6,252,916). We only have partial coverage of patents granted before 1976 (approximately 45%), and only information on their titles and abstracts, but not their claims. They are nonetheless included in order to establish a baseline dictionary.

For each patent, we concatenate title, abstract, and claims. Next, we lowercase the text and tokenize it to words using the following regular expression: [a-z0-9][a-z0-9-]*[a-z0-9]+|[a-z0-9]. We consider a word as a sequence of letters and numbers that could be separated by hyphens ("-"). Next, we remove words composed only by numbers, one-character words, stop words from the Natural Language Toolkit (NLTK) in the Python library[1], and words appearing in only one patent. In addition to natural stop words, we remove a manually compiled list of 32,255 very common keywords. First, we compile a list with the most frequently occurring keywords in patents. Next, we identify and exclude those keywords that are unrelated to the technical content of patents, but keep the frequently occurring technical keywords (e.g. internet, bluetooth, dna, rfid, glyphosate). The excluded keywords include both very common non-technical keywords (e.g. invention, discovery, claim, disclose, describe, include, patent) and very common mistakenly combined words which result from the tokenization process (e.g. comprisinga, combinefirst).

Next, we apply stemming to each word using the SnowBall method from the NLTK library. What remains is a collection of stemmed keywords which represent the technical content of the patent. The entire cleaned vocabulary contains 1,362,971 unique keywords and the average and median number of unique keywords per patent granted since 1976 is 61 and 56 respectively (stdev=29).

Open access data files

The txt file **"patent text raw"** contains one row and four columns for each U.S. utility patent granted between March 1969 and May 2018 (n=6,252,916). The first column contains the patent number. The second column contains the unprocessed/raw title text, the third column contains the raw abstract, and the fourth column contains the raw claims. Notice that we have no full coverage of patents granted before 1976, and only their titles and abstracts (no patent claims).

[1] Examples of stop words from the NLTK library include: the, am, been, does, for, has.

The txt file **"keywords"** contains one row and two columns for each U.S. utility patent granted between March 1969 and May 2018 (n=6,252,916). The first column contains the patent number. The second column contains the set of unique cleaned and stemmed keywords retrieved from the tittle, abstract, and claims of the patent. Notice that we have no full coverage of patents granted before 1976, and only their titles and abstracts (no patent claims). Words composed of numbers only, one-digit words, words which appear in only one patent, stop words, and frequently occurring non-technical words are removed. The keywords are separated by a single space and alphabetically ordered. This data can be used to measure and map the similarity between patents, inventors, firms, or geographical regions in technology space (e.g. Arts et al., 2018). The data can also be used to identify all patents related to a certain technology, to trace follow-on invention and diffusion of certain technologies or patents (e.g. de Rassenfosse et al., 2020), to measure knowledge spillovers from R&D (e.g. Myers and Lanahan 2020), or spillovers between science and technology (e.g. Iaria et al., 2018). One fruitful opportunity is to match the dataset with other text documents such as scientific publications (e.g. Iaria et al., 2018), funding opportunity announcements (e.g. Myers and Lanahan 2020), specifications of technical standards (Brachtendorf et al., 2020), or product descriptions (e.g. Argente et al. 2020). Another opportunity is to link this dataset with the datasets "new_keywords" and "new_keyword_comb_all" (see description below) in order to identify the patents that reuse a certain new keyword or keyword combination.

The txt file **"100_most_similar_patents"** contains for each U.S. utility patent granted between March 1969 and May 2018 up to 100 of the most similar patents based on cosine similarity from the entire population. To calculate cosine similarity, each patent is represented as a vector of 1,362,971 dimensions where each dimension corresponds to one keyword from the entire vocabulary and its value captures the frequency of this keyword in the particular patent document. For each patent, the file contains up to 100 rows and three columns (n=625,290,533). The first column contains the patent number. The second column contains the patent number of one of the 100 most similar patents (ordered by cosine similarity), and the third column contains the cosine similarity between the two patents. This file can be used for instance to identify related prior art, to assess the novelty of a patent, or to select a matched control group of similar patents (e.g. Arts et al., 2018).

The txt file **"1000_most_similar_patents"** contains for each U.S. utility patent granted between March 1969 and May 2018 up to 1,000 of the most similar patents based on cosine similarity from the entire population. For each patent, the file contains up to 1,000 rows and three columns (n= 6,252,899,662). The first column contains the patent number. The second column contains the patent number of one of the 1,000 most similar patents (ordered by cosine similarity), and the third column contains the cosine similarity between the two patents.

The txt file **"new_bigrams"** contains one row and four columns for each new bigram (two consecutive words in the patent document) introduced for the first time in history by granted U.S. utility patents filed between 1980 and February 2018 (n=7,128,180). Patents filed before 1980 are used to establish a baseline dictionary of bigrams. The first two columns contain the two keywords part of the new bigram. The third column contains the number of the patent introducing the bigram for the first time based on filing date. The fourth column counts the total number of patents which use this bigram in their title, abstract, or claims.

The txt file **"new_trigrams"** contains one row and five columns for each new trigram (three consecutive words in the patent document) introduced for the first time in history by granted U.S.

utility patents filed between 1980 and February 2018 (n=11,119,812). Patents filed before 1980 are used to establish a baseline dictionary of trigrams. The first three columns contain the three keywords part of the new trigram. The fourth column contains the number of the patent introducing the trigram for the first time based on filing date. The fifth column counts the total number of patents which use this trigram in their title, abstract, or claims.

The txt file **"new_keyword_comb_all"** contains one row and four columns for each new pairwise combination of keywords introduced for the first time in history by granted U.S. utility patents filed between 1980 and February 2018 (n= 670,585,951). To do so, we calculate all possible pairs between any of the keywords of a patent. In contrast to bigrams and trigrams, it does not matter where the keywords appear in the patent, nor the order in which they appear. Patents introducing a new keyword are a subset of patents with new keyword pairs because new keywords by definition result in new combinations. Patents filed before 1980 are used to establish a baseline dictionary of keyword combinations. The first column contains the first keyword part of the new keyword pair. The second column contains the second keyword part of the new keyword pair. Notice that it does not matter where the keywords appear in the title, abstract, or claims, neither the order in which they appear. The third column contains the number of the patent introducing the keyword pair for the first time based on filing date. The fourth column counts the total number of patents which use this keyword pair. Because the uncompressed file "new_keyword_comb_all" is approximately 20 GB, we also upload 6 segmented files (the years refer to the filing year):
"new_keyword_comb_1980_1989" (n=95,966,135),
"new_keyword_comb_1990_1994" (n=68,592,333),
"new_keyword_comb_1995_1999" (n=112,407,794),
"new_keyword_comb_2000_2004" (n=147,782,061),
"new_keyword_comb_2005_2009" (n=136,234,436),
"new_keyword_comb_2010_2018" (n=109,603,192).

The txt file **"cosine_similarity"** contains one row and three columns for each U.S. utility patent granted between March 1969 and May 2018 (n=6,252,916). The first column contains the patent number. The second column is the average cosine similarity between the focal patent and all patents filed in the five years before the focal patent based on the keywords retrieved from the tittle, abstract, and claims of the patent. The third column is the average cosine similarity between the focal patent and all patents filed in the five years after the focal patent. To calculate cosines, each patent is represented as a vector of 1,362,971dimensions where each dimension corresponds to one keyword from the entire vocabulary and its value captures the frequency of this keyword in the particular patent document. In contrast to *new_word, new_bigram, new_trigram,* and *new_word_comb* which isolate n-grams or keyword pairs, cosine similarity relies on the entire combination of keywords of a patent and also takes into account the frequency of a keyword in a patent.

The txt file **"patent_text_measures"** contains one row and eleven columns for each granted U.S. utility patent filed between 1980 and February 2018 (n=5,645,845). The first column contains the patent number. The second column contains the number of new keywords (unigrams) introduced by the patent (*new_word*). Filing dates are used to determine the first patent introducing a keyword. The third column contains the number of new keywords introduced by the patent weighted by their future reuse (*new_word_reuse*). For patent *p*, $new\_word\_reuse_p = \sum_{i=1}^{n}(1 + u_i)$ with *n* equal to the number of new keywords introduced by patent *p* and u_i equal to the number of future patents which reuse the new keyword *i*. The fourth column contains the number of new bigrams (two consecutive keywords in the patent document) introduced by the patent for the first time (*new_bigram*). In line

with the calculation of *new_word_reuse*, the fifth column contains the number of new bigrams introduced by the patent weighted by their future reuse (*new_bigram_reuse*). The sixth column contains the number of new trigrams (three consecutive keywords in the patent document) introduced by the patent (*new_trigram*). The seventh column contains the number of new trigrams introduced by the patent weighted by their future reuse (*new_trigram_reuse*). The eighth column contains the number of new pairwise keyword combinations introduced by the patent (*new_word_comb*). The ninth column contains the number of new pairwise keyword combinations introduced by the patent weighted by their future reuse (*new_word_comb_reuse*). The tenth column contains the average cosine similarity between the focal patent and all other patents filed in the five years before the focal patent (*backward_cosine*). The eleventh column contains the average cosine similarity between the focal patent and all other patents filed in the five years after the focal patent (*forward_cosine*).

References

Argente, D., Baslandze, S., Hanley, D., & Moreira, S. (2020). Patents to Products: Product Innovation and Firm Dynamics. FRB Atlanta Working Paper No. 2020-4, Available at SSRN: https://ssrn.com/abstract=3587377 or http://dx.doi.org/10.29338/wp2020-04

Arts, S., Hou, J., & Gomez, JC. (2020). Natural language processing to identify the creation and impact of new technologies in patent text: code, data, and new measures. Forthcoming *Research Policy*. (https://doi.org/10.1016/j.respol.2020.104144)

Arts, S., Cassiman, B., Gomez, J.C., (2018). Text matching to measure patent similarity. *Strategic Management Journal*, 39(1), 62-84.

Brachtendorf, L., Gaessler, F., & Harhoff, D. (2020). Truly Standard-Essential Patents? A Semantics-Based Analysis. CEPR discussion paper DP14726.

de Rassenfosse, G., Pellegrino, G., & Raiteri, E. (2020). Do patents enable disclosure? Evidence from the invention secrecy act. Available at SSRN: https://ssrn.com/abstract=3561896 or http://dx.doi.org/10.2139/ssrn.3561896

Iaria, A., Schwarz, C., & Waldinger, F. (2018). Frontier knowledge and scientific production: Evidence from the collapse of international science. *The Quarterly Journal of Economics*, 133(2), 927-991.

Marco, A.C., Sarnoff, J.D., deGrazia, C.A. (2016). Patent claims and patent scope, USPTO Economic Working Paper No. 2016-04.

Myers, K., & Lanahan, L. (2020). Research subsidy spillovers, two ways (June 9, 2020). Available at SSRN: https://ssrn.com/abstract=3550479 or http://dx.doi.org/10.2139/ssrn.3550479