US Patent Data Analysis: Impact of COVID

Data Science Institute
COLUMBIA UNIVERSITY

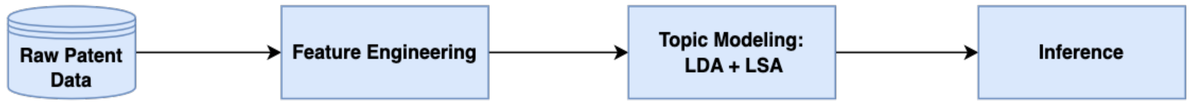
Authors: Arunit Maity, Malaika Gupta, Mehrab Singh Gill, Sanjeev Tewani, Sarthak Bhargava, Shreya Verma
Faculty Mentor: Dr. Jorge Guzman

Data Science Capstone Project with Dr. Jorge Guzman, Columbia Business School

Introduction

The US patent application procedure safeguards the rights of innovators and gives them an incentive to transform their novel ideas into beneficial products. Filing for a patent is a crucial first step in this process of innovation. We think that both the method of innovation and the direction of innovative activities may have been significantly impacted by the COVID-19 pandemic. Therefore, our project's objective is to gauge (i) pace of innovations, (ii) nature of innovations and (iii) examine changes in the location.

Methodology



- Scrape from USPTO
- Upload to AWS S3 for easy access

- Split XML files into individual applications
- Extract important features
- Lemmatize and remove stop words
- Construct bag of words from claims

- TFIDF
- Hyperparameter tuning
- Coherence score: UMass metric
- Obtain optimal number of clusters for each year

- Analysis of statistical significance across years
- Identification of relevant topics

Figure 1. Process Workflow

Visualization and Understanding the Evolution of Location

To obtain a better understanding of the dataset and its distribution, we visualized and analyzed the trends with respect to parameters such as patent counts, number of inventors, organization city and organization country. This is depicted in the plots as below.

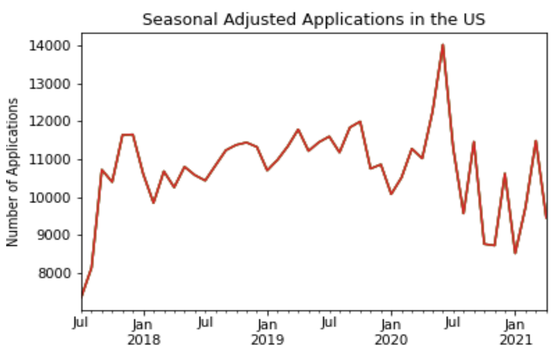


Figure 2. Seasonally adjusted patent applications in the US across years

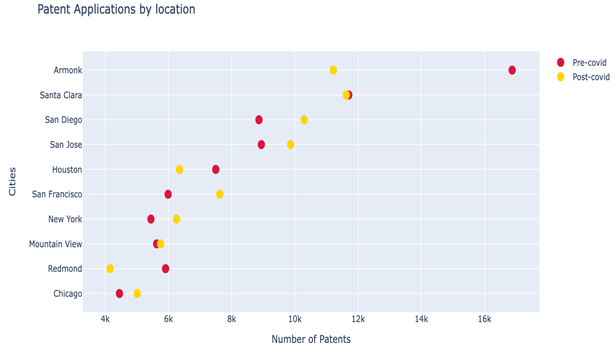


Figure 3. Patent Applications Pre and Post COVID

Looking at Figure 2, we see a clear spike in patent filing in June 2020, and around the same time the impact of COVID was being felt around the globe. Additionally, it is evident from Figure 3 that in most cities, with the highest patent applications, there is an increase in patent filing after COVID.

Results

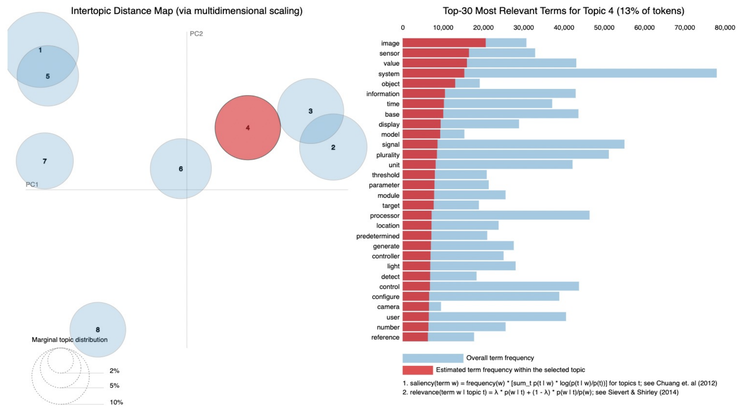
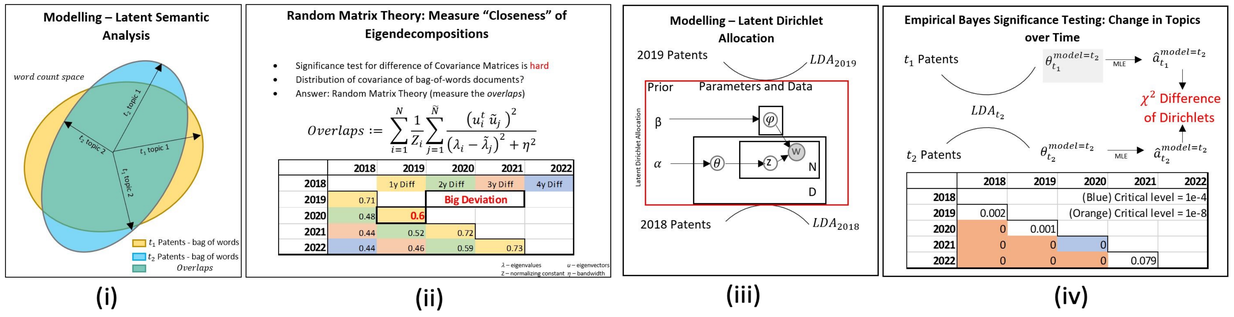


Figure 4. LDA Visualization for 2018

From LDA's inter-topic Distance plot shown on the left in Figure 4, we identify how these eight clusters relate to each other, including potential higher-level structure between groups of topics.
The bar graph in Figure 4 displays the top frequent or relevant terms within each topic.



(i) (ii) (iii) (iv)

Figure 5. Measure of significance using LSA (i, ii) and LDA (iii, iv)

The measure of significance for LSA quantifies the eigen decomposition overlaps between years which shows that 2019-2020 have a uniquely low overlap in eigenspaces. Similarly, for LDA, empirical bayes MLE estimate of the Dirichlet parameters is computed by running held-out years through the target year's model. The test shows that the topic distribution of 2021 was significantly different from 2020.

Conclusion

Leveraging two different topic modeling approaches and their underlying distributional assumptions, this study shows that the pandemic period did show a significant change in the distribution of US patents. It could be concluded that the COVID-19 pandemic may have meaningfully altered the focus of innovative efforts.

Acknowledgements

We appreciate the support of Dr. Jorge Guzman for his guidance throughout the project.

References

- Bun, J., Bouchaud, J. P., & Potters, M. (2018). Overlaps between eigenvectors of correlated random matrices. *Physical Review E*, *98*(5), 052145.
- Minka, T. (2000). Estimating a Dirichlet distribution.