Patent Data Analysis

Abhishek Jha, Boyan Tian, Matthew Cooper, Zifan Wang
CS 4624 Multimedia, Hypertext and Information Access, Dr. Edward A. Fox
Virginia Tech, Blacksburg VA 24061, 27 April 2021
Presentation Overview

• Project Recap
• Final Design
• Term Frequency-Inverse Document Frequency
• The Road Ahead
• Lessons Learned
• Acknowledgements
Project Recap

• Three parts of data analytics
  • Where do we fit in?

Deliver a Jupyter Notebook that scrapes Google Patents.
Term Frequency-Inverse Document Frequency (TF-IDF)

- What is TF?
- What is IDF?
- Why use them?
- Allows classification based on the terms used in the patent itself.
The Road Ahead

2 - 16 Feb 2021
- Sign the contract
- Select packages and tools for the scripts.
- Presentation 1
- Begin development

17 Feb - 15 Mar 2021
- Script V1.0
- Testing Output
- Selection of Analysis
- Begin V1.1

16 Mar - 6 Apr 2021
- Script V1.1
- Text Preprocessing
- Methodology
- Presentation 2
- Interim Report

6 Apr - 4 May 2021
- Documented Code
- Data Analytics
- Final Report
- User Manual
- Final Presentation

5 May 2021
- All deliverables
- Enjoy Summer
Lessons Learned

- Constant communication is essential
- Starting Early
  - Baby step execution
  - Keeping track of deadlines
- We should buy an Agile Development handbook
Acknowledgments

Dr. Florian Zach
Howard Feiertag Department of Hospitality and Tourism Management
florian@vt.edu
Questions...