MIE 1624 Introduction to Data Science and Analytics – Fall 2023

Course Project

Course Instructor: Dr. Oleksandr Romanko
Teaching Assistants: Sina Davari, Arnaud Deza, Eric Floro, Minori Narita

Deadline: Sunday, December 3, 11:59pm

Background

Global political crises, like russia's war against Ukraine, underscore the necessity for global innovations to accelerate economic growth, enhance competitiveness, and empower nations to resist aggressions. We need to review the state of the Canadian innovation ecosystem and develop a new strategy to facilitate innovations in Canada. In 2017, Innovation, Science and Economic Development Canada (ISED) announced the "Innovation for a better Canada" strategy https://ised-isde.canada.ca/site/innovation-better-canada. In 2019, they published a 100-page report "Building a Nation of Innovators". The report did not take into account new challenges (COVID-19 pandemic, war in Ukraine) and new opportunities (development of Generative AI models such as ChatGPT) that appeared since the report publication in 2019.

Your team has been hired *pro bono* by the University of Toronto as a consultant to evaluate how Canada measures among other countries in terms of innovations, rewrite or enhance the current strategy draft, and propose practical steps to implement a proposed strategy. Your task is to analyze various traditional media, social media, reports, YouTube videos and other sources of information in different countries to determine how Canada can accelerate innovations. You need to write a consulting report and deliver a presentation presenting your findings and your suggestions to the Government of Canada for changes in its current portfolio of strategies that are projected to have a positive impact on Canada's economic development and help Canada to secure peace and prosperity among other nations in the future.

Data science, analytics, AI, and big data are becoming widely used in many fields, that leads to the ever-increasing demand for data analysts, data scientists, managers of analytics and AI, and other data professionals. Due to that, text analytics and natural language processing is now a hot topic for researchers, educators, and entrepreneurs. In addition, the design of questionnaires and surveying combined with statistical analysis and trend building and projections, allows us to perform advanced data analytics studies. The development of Generative AI models, such as ChatGPT, gives an opportunity to summarize text information and search for insights and ideas in large text datasets.

In this project, you need to perform an analysis of key-factors responsible for differences in innovation across countries. To do this your team will *first* need to collect public data from news, social media, reports, blogs, and interviews with experts about the development of innovative ecosystems in different countries and their innovation strategies. You need to summarize how innovations are measured in different countries and how countries are rated relative to each other. You can use the Global Innovation Index 2023 report, published annually by the World Intellectual Property Organization at https://www.globalinnovationindex.org, as a primary source for such a comparison. For instance, one possible measure of innovations can be a number of unicorns (startups valued at $1B or more) in each country. You need to analyze which factors (e.g., human capital, productivity, R&D spendings, research tax breaks, etc.) drive innovations, how Canada stands on those factors, and which factors can be critical to improving Canada's position relative to other countries.

Second, your team will need to come up with an enhanced proposal about Canada's ecosystem development strategy. To do this you will be using generative AI models to summarize experiences of other countries from reports, strategy papers, scientific publications, social media (especially Twitter/X), expert opinions. Adjusting those strategies for a situation in Canada is also important. You can use the report published by ISED as a starting point for developing your strategy, shorten and enhance it. Alternatively, you may develop your strategy from scratch. Materials collected by the course instructor are available for you at https://drive.google.com/drive/folders/1TbCTw-W2lhQXgbN0iuFvIcLZViA0PAon?usp=sharing (Google Drive folder).

Finally, based on the results of your analysis, your team will need to develop data-driven practical steps (government programs, internship programs, industry-academic collaboration initiatives, international collaborations and collaborations with Canadians abroad, partnerships with multinational companies, tax incentives for R&D projects, defense funding guidelines, funding from international organizations and agencies, etc.) to implement the proposed strategy in practice. Your practical steps may be tied up with a PR strategy to promote Canada in media and social media as "Canada – the country of innovations". You may want to develop visualizations (e.g., infographics), key messages, and propose news stories to perform storytelling. The book "Ingenious: How Canadian Innovators Made the World Smarter, Smaller, Kinder, Safer, Healthier, Wealthier, and Happier" may give you some ideas related to storytelling for promoting Canadian innovation.

Learning Objectives

- Develop the ability to work in a team on a consulting project. (You are required to work on the project in the same group as for your in-class presentation. Check the Quercus portal for the list of your group members.)

- Improve on skills and competencies required for performing a full cycle of data science and analytics workflow, i.e., data collection and pre-processing, applying algorithms to analyze data, trend identification, storytelling based on analytics (writing a consulting report and delivering an oral presentation). This is an open-ended problem, as such creative solutions are encouraged where possible.

Tools Allowed

- You can use Python libraries mentioned in-class as well as any other Python libraries you find during your research. Note that you can only use Python 3.
- If you plan to use other tools beyond Python, including but not limited to software products, cloud services, chatbots, during your research, save outputs of those to file(s) and read to your Python code. You should not be restricted by Python limitations in this project.
- For visualizing results in your report and presentation you may use Python or any other outside tool, e.g., Excel, Tableau, Power BI, etc.

To Do Summary

Perform analytics on news articles, reports, social media posts, books, YouTube videos, search results, etc., to find out how Canada can innovate better. You are not restricted in the sources of information that you can analyze, e.g., New York Times, Thomson Reuters, Associated Press, Facebook, Twitter/X, Instagram images, Amazon, Google search, and in the manner you get access to that information, e.g., APIs or web-scraping. (You may want to perform automatic or manual fact-checking of the information that you analyze in order to assess your data's reliability.) Use of Generative AI tools, such as ChatGPT, is encouraged. You may feed analyst reports and other relevant documents as PDF/DOCX files to Generative AI models (GPT-3.5, GPT-4, Bard AI, Claude 2, etc.) to summarize documents and to extract ideas for your strategy document. You may classify or cluster information about different topics related to innovations in Canada and other countries, and then try to cover all the identified topics or concentrate only on some of those. You can use the algorithms covered in the course as well as any other algorithms that you find during your research. (Feel free to use cloud services on Amazon AWS, Google Cloud, Microsoft Azure, IBM Cloud for your analysis or data retrieval.)

Among others, you may consider utilizing the following algorithms for this project:

- Sentiment analysis of social media posts and news articles related to innovations.
- Factor and topics identification, e.g., based on your sentiment analysis results you may try to identify factors/reasons/topics that drive sentiment.
- Clustering, e.g., clustering of practical steps and governmental programs that support innovations in different countries.
- Optimization models, e.g., optimizing government funding and budgets to support innovations.
- Trend building with regression models.
- Classification algorithms.
- Generative AI algorithms.

Note: the scope of the project is quite wide, and it is advised that you narrow it down based on your interests and expertise. Make the project truly yours.

TO DO

Finish the following four parts **based on your data science and AI-based analytics:**

Part 1 – Collect data and summarize how innovations are measured:

Collect public data from news, social media, reports, blogs, interviews with experts about development of innovation ecosystems in different countries and their innovation strategies. You need to summarize how innovations are measured in different countries and how countries are rated relative to each other. You can use Global Innovation Index 2023 report, published annually by the World Intellectual Property Organization at https://www.globalinnovationindex.org, as a primary source of such comparison. For instance, one possible measure of innovations can be a number of unicorns (startups valued at $1B or more) in each country. You need to analyze which factors (e.g., human capital, productivity, R&D spendings, research tax breaks, etc.) drive innovations, how Canada stands on those factors, and which factors can be critical to improve Canada's position relative to other countries.

If you find out that the dataset provided by the course instructor in the Google Drive folder at https://drive.google.com/drive/folders/1TbCTw-W2lhQXgbN0iuFvIcLZViA0PAon is not enough for your purposes, feel free to use additional datasets available on the web. Among other datasets, you may consider:

- *Opinion of influencers* – tweets of people that discuss innovations (you may use Twitter/X API from Python to collect tweets related to innovations);
- *Social media analysis* (sentiment, topic, innovation measures);
- *Kaggle datasets collected by people*, e.g., https://www.kaggle.com/datasets/saurabhshahane/digital-innovation;
- *YouTube videos*, e.g., https://www.youtube.com/watch?v=WNTtWYFhWus;
- *Data World and other "open data" datasets* https://data.world/datasets/innovation;
- *News aggregators* (API or web-scraping), e.g. https://www.pressreader.com/search?query=innovation&orderBy=Relevance&searchFor=Articles;
- *Google Trends* https://trends.google.com, Kaggle, and similar data/news aggregation portals.

Feel free to experiment with Python libraries and APIs to download news articles, e.g., https://newsapi.org/docs/client-libraries/python, and Python libraries such as `PyPDF2` and `PyMuPDF` to convert PDF files to text.

Summarize results from your datasets, e.g., with Generative AI models (GPT-3.5, GPT-4, Bard AI, Claude 2) using APIs from Python, or user interfaces such as https://chat.openai.com. For each dataset, make conclusions about your results. Explain those. Select summaries of datasets that you plan to use in Part 2.

Part 2 – Develop proposal for Canada’s Innovation Ecosystem Development Strategy:

Your team will need to come up with an enhanced proposal about Canada’s ecosystem development strategy. You can use the report published by ISED as a starting point for developing your strategy, shorten and enhance it. Alternatively, you may develop your strategy from scratch. You are not restricted in how you structure your report and presentation, but possible sections of the strategy proposal should include results of Part 1, Part 2, Part 3, and Part 4.

Summarizing experiences of other countries, that you performed in Part 1 with generative AI models from reports, strategy papers, scientific publications, social media (especially Twitter/X), expert opinions, may be a key point in your analysis. Adjusting those strategies for a situation in Canada is also important.

Summarize results of your analysis in the form of a consulting report and presentation slides.

Part 3 – Practical steps for implementing your strategy:

Based on your analysis, your team will need to develop practical steps (government programs, internship programs, industry-academic collaboration initiatives, international collaborations and collaborations with Canadians abroad, partnerships with multinational companies, tax incentives for R&D projects, defense funding guidelines, funding from international organizations and agencies, etc.) to implement your proposed strategy in practice. As an example of a practical step, you may analyze Canada’s Mitacs research internship program and propose enhancements or modifications of it. Your strategy and practical steps may be tied up with a PR strategy to promote Canada in media and social media as “Canada – the country of innovations”, that you perform in Part 4.

Part 4 – Visualizations, storytelling, recommendations:

Your task is to visualize modeling results obtained in Part 1, 2 and 3. Design visualization(s), e.g., wordclouds, that allows decision makers to grasp the current state of innovation ecosystem development in Canada, key findings and milestones of your proposal, practical implementation steps that you propose, etc. Your strategy and practical steps should be tied up with a PR strategy to promote Canada in media and social media as “Canada – the country of innovations”. You may want to develop visualizations (e.g., infographics), key messages, and propose news stories to perform storytelling based on your analysis.

In addition, based on your analysis and identification of key factors/reasons/topics/steps you need to develop a narrative (storytelling via presentation and report) presenting your findings and your suggestions to the Government of Canada, and national and international NGOs for applying for funding to support the portfolio of strategies that your team developed. Your storytelling should concentrate to have a projected positive impact on Canada’s international presence and image. If necessary, you may compliment your analysis with posts/tweets templates that you propose to publicize your strategy.

Note: the scope of the questions in Part 1-4 is quite wide, and it is advised that you narrow it down based on your interests and expertise. Make the work truly yours.

Project Presentations

- Project presentations are scheduled for **Tuesday, December 5, 6:00-9:00pm (room SF 1101)**, doors are open to the public.
- **Do not make your presentation overly technical**. Your audience is business-oriented and may know little about data science, people are interested in the insights that you got from your analysis and why your results can and should be used for decision-making.

What to Submit via Quercus

1. Your Jupyter notebook with appropriate documentation for every step as well as the relevant data files. Comment out any data retrieval processes (e.g., from web scraping, downloading, APIs, etc.) in your code and replace it with code for reading the corresponding data from files. If you decide to do some of your analysis outside of Python, save your results to data file(s) and read those into your Jupyter notebook. (**Submit all those data files together with your Jupyter notebook**). Consider the Jupyter notebook as what you would report to senior data scientists and machine learning engineers. Documentation and comments in the Jupyter notebook should contain technical details of your data analysis and be understandable by data science professionals if they run it on their own. Make sure that your Jupyter notebook runs on Google Colab https://colab.research.google.com portal and that all needed data files are included in your submission. If the size of the data files exceeds Quercus's capacity, those should be stored on a cloud drive (e.g., Dropbox, Google Drive), and the link to the directory should be included in the notebook.
2. A 5 to 10-page consulting report in PDF and DOCX formats that summarizes your findings and results (all graphs should have axes appropriately labelled, all visual materials should be understandable and the graphics of sufficient quality to be easily readable.) This report should be business oriented and cover your problem more extensively than your presentation.
3. Your business-oriented presentation slides in PowerPoint and PDF formats. (Each group will present their findings and results during an in-class 7-minute presentation with 1-2 minutes for questions. Presentations will be timed and stopped after 7 minutes.) It is up to you how many group members are presenting (one, two or all).
4. Video recording of your 7~8-minute presentation in mp4 or avi format. You do not need a video recording of you, you're audio on-top of your slideshow presentation is fine.

Marking

- The **project is worth 20 points** (10 points for your business-oriented presentation and 10 points for your analysis and report).
- The **presentation** will be graded as follows (**10 total pts**):
  1. *Organization & Delivery* (**3 pts**)
     - Clear structure and logical flow, with an introduction, main content, and conclusion;
     - Demonstrated enthusiasm and poise throughout the presentation;
     - Effective time management ensuring all topics are covered without rushing within the given time frame.
  2. *Content* (**3 pts**)
     - Use of appropriate, impactful visuals that complement and enhance the spoken content;
     - Clear representation of high-level ideas, ensuring understanding for all audience levels;
     - Proactive and comprehensive answering of questions, showcasing depth of understanding.
  3. *Business Pitch* (**4 pts**)
     - Concrete recommendations backed by data and analysis;
     - Clear problem-solution correspondence, showing how the solution addresses the identified issue;
     - Tailored relevance to the target audience, ensuring the pitch resonates and is compelling.
- The **analysis** in **Jupyter notebook** and the **report** will be graded as follows (**10 total pts**):
  1. *Problem Definition & Data Search* (**2 pts**)
     - Skillful narrowing down of the broader problem to a focused research question;
     - Clearly defined problem statement that is concise yet comprehensive;
     - Efficient identification and sourcing of relevant data, ensuring data integrity.
  2. *Analysis* (**5 pts**)
     - Cleaning the data;
     - Relevant and insightful visualizations that aid understanding;
     - Proper application of algorithms, evaluation of algorithms' accuracy, and showcasing technical proficiency.
  3. *Discussion & Insights* (**3 pts**)
     - Clear and direct link between the analysis conducted and the problem identified;
     - Utilization of effective data-driven decision-making criteria, demonstrating critical thinking;
     - Engaging storytelling that makes the data and analysis relatable and memorable.
- Every group member gets the same mark for the project. It is your responsibility to determine how you split the work inside your group. At least half of your group needs to be present during the in-class project presentations to answer questions.

<u>Notes</u>

- For the deliverables, consider the Jupyter notebook as what you would report to senior data scientists and machine learning engineers, and the consulting report and the presentation as what you would report to PR managers, university officials, government officials, journalists, and general audience.
- The presentation would be a visual representation of the executive summary of your report.
- The audience for your presentation and report in particular is business-oriented and includes people who are interested in the insights you gathered from your analysis and how your results should be used for decision-making.