Master thesis proposal

Title:

*Statistical analysis of Towards Data Science articles:*

*analyzing reader preferences for theoretical vs. technological content*

1. Background

Towards Data Science (TDS) is a **widely recognized online platform** on Medium, **dedicated to sharing high quality content about data science**, machine learning, artificial intelligence, and related fields. The articles are written in a popular science style but lean more toward academic publications than journalistic writing.

It is aimed at both enthusiasts of data science and related fields, as well as programmers, professionals, and industry experts. A typical article includes explanations of concepts, solutions, algorithm descriptions, source code, and usage recommendations.

The extensive collection of articles covers the latest discoveries and trends, tutorials on using programming libraries, and introductions to the mathematical and statistical foundations of data science, machine learning, and AI.

The question that the writer must ask themselves is, of course**, what topic is worth writing about in order for the article to be useful to as many readers as possible**. Should it be a trendy or niche topic, focused on the technical aspects of a solution or the theoretical foundations of methods, a single article or a series on a specific topic, etc.

1. Objectives

There is a category of articles that explore the theoretical foundations (mathematical or statistical) of algorithms and tools in data science, machine learning, and AI. These articles differ from those focused on technology and programming in that they:

* have a broader scope,
* are typically aimed at a different audience,
* and may have a longer lifespan, maintaining reader interest over time compared to articles on contemporary issues or the current state of technology.

The aim of this work is **to statistically describe the distinctive properties of** the aforementioned category of articles, which could be referred to as ***'theoretical background'*** (TH) ***articles*** in contrast to ***'****technology-centered****'*** (TECH)*articles*.

This requires collecting data on a significant number of such articles and performing **statistical** and **exploratory data analysis (EDA) on this collection**.

During the thesis project, **the following tasks should be carried out**.

1. Rigorous Statistical Analysis.
   1. Providing a precise **definitions** of a *'theoretical background'* and ***'****technology-centered****'*** articles, such that the search criteria can be derived from them.
   2. **Formulating hypotheses** regarding the characteristics of TH articles, such as 'TH articles maintain interest for a longer period compared to TECH articles,' or 'The distribution of views for TH articles over time has a broader spread than that of TECH articles,' and similar statements.
   3. Collecting **a representative sample** (dataset) suitable for testing the formulated hypotheses using web scraping techniques.
   4. Testing the hypotheses using statistical inference methods, including test design, reporting p-values or confidence intervals.
2. Exploratory analysis
   1. Selecting features that describe the articles.
   2. Using web scraping techniques to gather as much data as possible on THS articles (of any type, not necessarily TH or TECH) within the given time limit.
   3. Visual inspection of the dataset: heatmaps, crosstabs, distributions, chord diagrams, and so on.
   4. Identifying patterns in the data using machine learning techniques, such as clustering and assessing feature importance in regression models or classifiers.
   5. Using clustering techniques to identify natural groups of articles and verify whether the previous assumption about the existence of distinct TH and TECH groups was justified, possibly detecting other groups worth examining.
   6. Reporting detected patterns.
3. Data

The data must be **up-to-date** and should cover at least back to 2018 (TDS started in 2016). A program should be developed to fetch article data using web scraping techniques.

In addition to general-purpose libraries like Beautiful Soup:

<https://hackernoon.com/how-to-scrape-a-medium-publication-a-python-tutorial-for-beginners-o8u3t69>

there are tools specifically designed for the Medium platform, where TDS articles are hosted:

* <https://axiom.ai/scrape/medium>
* <https://www.octoparse.com/blog/how-to-scrape-medium-data>
* <https://medium.com/free-code-camp/how-i-analyzed-a-million-medium-articles-to-learn-how-many-claps-great-stories-get-d688cf92759f>
* <https://github.com/harrisonjansma/Medium_Scraper>
* <https://www.browse.ai/t/extract-stories-medium-user-profile>

1. Required background

Good programming skills.

Solid knowledge of statistical inference methods.

1. Contact person

Jarosław Drapała

Faculty of Information and Communication Technology

Wrocław University of Science and Technology

e-mail: [jaroslaw.drapala@pwr.edu.pl](mailto:jaroslaw.drapala@pwr.edu.pl)