Well-being index prediction from Google search data

Aurélien Bellet, Pascal Denis

October 15, 2018

Supervisors

- Team Magnet, INRIA/CRIStAL: http://team.inria.fr/magnet
- Aurélien Bellet (aurelien.bellet@inria.fr), Pascal Denis (pascal.denis@inria.fr)

Keywords

Machine Learning, Natural Language Processing, Word Embeddings, Political Science.

Context

The well-being of populations at the national or regional level can be used to evaluate a society's health and the effectiveness of public policies beyond traditional measures like such as GDP growth. Such well-being measure is typically obtained through polls such as those run by Gallup[1] on a daily basis in the United States US and yearly for 155 countries. These polls are expensive to run, and may not reflect all aspects of well-being. As an alternative to surveys ("stated well-being"), one can study the actual decisions that people make ("revealed well-being"). This motivates measuring well-being from other sources of data. Notably, the work of [1] constructs a proxy measure by regressing the Gallup index using Google Internet search data (volume of queries across time).

Objectives

The work of [1] is based on a small number of manually selected search queries grouped into relevant categories (job search, financial security, healthy habits, family life, etc). We will provide some Python code to query the API of Google Trends[2] and to reproduce the experiments of [1].

[1] https://wellbeingindex.sharecare.com/
[2] https://trends.google.com/trends/

The goal of this project is to extend the above work to improve the prediction performance and/or remove the need for manual query and category selection. In particular, we will explore the following directions:

- Automatic selection of relevant queries by learning sparse prediction models.
- Use of multi-task learning to jointly predict multiple Gallup indices.
- Use of word embeddings [2] to automatically generate relevant queries from a small set of seed words.

References

[1] Y. Algan, E. Beasley, F. Guyot, K. Higa, F. Murtin, and C. Senik. Big Data Measures of Well-Being: Evidence From a Google Well-Being Index in the United States. Technical report, OECD Statistics Working Papers, 2016.

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. Technical report, arXiv:1301.3781, 2013.